



Journal of  
*Imaging*

# Fine Art Pattern Extraction and Recognition

Edited by

Fabio Bellavia, Giovanna Castellano and Gennaro Vessio

Printed Edition of the Special Issue Published in *Journal of Imaging*

[www.mdpi.com/journal/jimaging](http://www.mdpi.com/journal/jimaging)



# **Fine Art Pattern Extraction and Recognition**



# **Fine Art Pattern Extraction and Recognition**

Editors

**Fabio Bellavia**

**Giovanna Castellano**

**Gennaro Vessio**

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin



*Editors*

Fabio Bellavia Department of Math and Computer Science, University of Palermo Italy	Giovanna Castellano Department of Computer Science, University of Bari Italy	Gennaro Vessio Department of Computer Science, University of Bari Italy
---	---	--

*Editorial Office*

MDPI  
St. Alban-Anlage 66  
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Journal of Imaging* (ISSN 2313-433X) (available at: [https://www.mdpi.com/journal/jimaging/special\\_issues/faper2020](https://www.mdpi.com/journal/jimaging/special_issues/faper2020)).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name Year, Volume Number*, Page Range.

**ISBN 978-3-0365-2225-8 (Hbk)**

**ISBN 978-3-0365-2226-5 (PDF)**

© 2021 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license, which allows users to download, copy and build upon published articles, as long as the author and publisher are properly credited, which ensures maximum dissemination and a wider impact of our publications.

The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons license CC BY-NC-ND.

# Contents

<b>About the Editors . . . . .</b>	vii
<b>Fabio Bellavia, Giovanna Castellano and Gennaro Vessio</b>	
Editorial for Special Issue “Fine Art Pattern Extraction and Recognition”	
Reprinted from: <i>J. Imaging</i> 2021, 7, 195, doi:10.3390/jimaging7100195 . . . . .	1
<b>Claudia Daffara and Elisa Marini</b>	
A Portable Compact System for Laser Speckle Correlation Imaging of Artworks Using Projected Speckle Pattern	
Reprinted from: <i>J. Imaging</i> 2020, 6, 119, doi:10.3390/jimaging6110119 . . . . .	5
<b>Marco Trombini, Federica Ferraro, Emanuela Manfredi, Giovanni Petrillo and Silvana Dellepiane</b>	
Camera Color Correction for Cultural Heritage Preservation Based on Clustered Data	
Reprinted from: <i>J. Imaging</i> 2021, 7, 115, doi:10.3390/jimaging7070115 . . . . .	27
<b>Marco Fanfani, Carlo Colombo and Fabio Bellavia</b>	
Restoration and Enhancement of Historical Stereo Photos †	
Reprinted from: <i>J. Imaging</i> 2021, 7, 103, doi:10.3390/jimaging7070103 . . . . .	47
<b>Annamaria Amura, Alessandro Aldini, Stefano Pagnotta, Emanuele Salerno, Anna Tonazzini and Paolo Triolo</b>	
Analysis of Diagnostic Images of Artworks and Feature Extraction: Design of a Methodology	
Reprinted from: <i>J. Imaging</i> 2021, 7, 53, doi:10.3390/jimaging7030053 . . . . .	67
<b>Fabrizio Banfi and Alessandro Mandelli</b>	
Computer Vision Meets Image Processing and UAS PhotoGrammetric Data Integration: From HBIM to the eXtended Reality Project of Arco della Pace in Milan and Its Decorative Complexity	
Reprinted from: <i>J. Imaging</i> 2021, 7, 118, doi:10.3390/jimaging7070118 . . . . .	85
<b>Nicolò Oreste Pinciroli Vago, Federico Milani, Piero Fraternali and Ricardo da Silva Torres</b>	
Comparing CAM Algorithms for the Identification of Salient Image Features in Iconography Artwork Analysis	
Reprinted from: <i>J. Imaging</i> 2021, 7, 106, doi:10.3390/jimaging7070106 . . . . .	117
<b>Mridul Ghosh, Sk Md Obaidullah, Francesco Gherardini and Maria Zdimalova</b>	
Classification of Geometric Forms in Mosaics Using Deep Neural Network	
Reprinted from: <i>J. Imaging</i> 2021, 7, 149, doi:10.3390/jimaging7080149 . . . . .	139
<b>Tsegaye Misikir Tashu, Sakina Hajiyeva and Tomas Horvath</b>	
Multimodal Emotion Recognition from Art Using Sequential Co-Attention	
Reprinted from: <i>J. Imaging</i> 2021, 7, 157, doi:10.3390/jimaging7080157 . . . . .	151
<b>Eva Cetinic</b>	
Towards Generating and Evaluating Iconographic Image Captions of Artworks	
Reprinted from: <i>J. Imaging</i> 2021, 7, 123, doi:10.3390/jimaging7080123 . . . . .	163
<b>Yalemisew Abgaz, Renato Rocha Souza, Japesh Methuku, Gerda Koch and Amelie Dorn</b>	
A Methodology for Semantic Enrichment of Cultural Heritage Images Using Artificial Intelligence Technologies	
Reprinted from: <i>J. Imaging</i> 2021, 7, 121, doi:10.3390/jimaging7080121 . . . . .	179

Aline Sindel, Thomas Klinke, Andreas Maier and Vincent Christlein

# ChainLineNet: Deep-Learning-Based Segmentation and Parameterization of Chain Lines in Historical Prints

Reprinted from: *J. Imaging* 2021, 7, 120, doi:10.3390/jimaging7070120 . . . . . 201

## About the Editors

**Fabio Bellavia** is Assistant Professor at the Department of Math and Computer Science of the University of Palermo. His research interests include computer vision and image processing, focusing on local image detectors and descriptors, image matching, 3D reconstruction, image mosaicing and stitching, color correction, and their applications to autonomous driving, forensic science and cultural heritage. He actively collaborates with the CVG Lab of the University of Florence he joined from 2012 to 2019 and, more recently, with the Center for Machine Perception of Czech Technical University, Prague. His current position is founded by a European PON AIM project on “Computational Methods for Cultural Heritage” and he is currently working on 3D reconstruction applications for the preservation of archaeological sites in collaboration with the Parco Archeologico e Paesaggistico della Valle dei Templi di Agrigento. He has co-authored several papers, which appeared in both international journals and conferences, such as IEEE TPAMI, IJCV and IEEE TIP. He is Associate Editor of IET Image Processing and serves as a reviewer for many international journals. He is also part of the program committee of several international conferences.

**Giovanna Castellano** is Associate Professor at the Department of Computer Science of the University of Bari and coordinator of the Computational Intelligence Laboratory. Her research interests are in the area of computational intelligence and include fuzzy image processing and computer vision, fuzzy systems, fuzzy clustering, image processing, image retrieval, neural networks, neuro-fuzzy modeling, granular computing and recommender systems. She is co-author of over 200 papers in peer-reviewed books, conference proceedings and international journals covering the above topics. She is Associate Editor of Information Sciences, Evolving Systems and International Journal of Intelligent Systems. She is member of the Program Committee of several refereed conferences. She acts as a reviewer for several international scientific journals published by high-rank publishers (Elsevier, IEEE, Springer) and for international conferences.

**Gennaro Vessio** is currently Assistant Professor at the Department of Computer Science of the University of Bari. His position is currently founded by a European PON AIM project on “Computational Methods for Cultural Heritage”. He is involved in the research activities of the Computational Intelligence Laboratory, coordinated by Prof. Giovanna Castellano. His current research interests include pattern recognition, machine and deep learning, and computer vision, and their application to diverse domains, including biometrics, e-health and digital humanities. He regularly serves as a reviewer for many international journals published by high-level publishers, including Elsevier, IEEE and Springer, and as a member of the Program Committee of many international conferences, such as SEKE.





Editorial

# Editorial for Special Issue “Fine Art Pattern Extraction and Recognition”

Fabio Bellavia <sup>1</sup>, Giovanna Castellano <sup>2</sup> and Gennaro Vessio <sup>2,\*</sup>

<sup>1</sup> Department of Math and Computer Science, University of Palermo, 90133 Palermo, Italy;  
fabio.bellavia@unipa.it

<sup>2</sup> Department of Computer Science, University of Bari, 70125 Bari, Italy; giovanna.castellano@uniba.it

\* Correspondence: gennaro.vessio@uniba.it

Cultural heritage, especially the fine arts, plays an invaluable role in the cultural, historical, and economic growth of our societies. Works of fine arts are primarily developed for aesthetic purposes and are mainly expressed through painting, sculpture, and architecture. In recent years, owing to technological improvements and drastic cost reductions, a large-scale digitization effort has been made, which has led to the increasing availability of large, digitized fine art collections. Coupled with recent advances in pattern recognition and computer vision, this availability has provided, especially researchers in these fields, with new opportunities to assist the art community by using automatic tools to further analyze and understand works of fine arts. Among other benefits, a deeper understanding of the fine arts has the potential to make works more accessible to a wider population, both in terms of fruition and creation, thus supporting the spread of culture.

The call for papers for the Special Issue “Fine Art Pattern Extraction and Recognition” was opened to anyone wishing to present advancements in the state of the art, innovative research, ongoing projects, and academic and industrial reports on the application of visual pattern extraction and recognition, for a better understanding and fruition of works of fine arts. The Special Issue solicited contributions from researchers in diverse areas such as pattern recognition, computer vision, artificial intelligence, and image processing. Furthermore, we also solicited the submission of papers as an extension of the works presented at the homonymous workshop we organized at the 25th International Conference on Pattern Recognition [1].

The Special Issue received several submissions, which underwent a rigorous peer review process. After the review process, 11 articles were selected based on the ratings and comments. The published articles cover various applications of cultural heritage and digital humanities research; focus on different branches fine arts such as painting, architecture, and photography; and develop and apply a range of techniques, from image processing to computer vision, based on handcrafted features and deep learning.

Artworks are subject to alterations over time due to various factors such as natural aging, external agents, inadequate conservation treatments, etc. Hence, techniques for monitoring, preserving, and restoring cultural heritage artifacts have become crucial. To this end, Daffara and Marini [2] present a non-destructive examination tool based on laser interferometry that uses laser speckle imaging for the effective mapping of subsurface defects in paintings. The system has been designed to be flexible, able to optimize its performance through an easy parameter adjustment. Trombini et al. [3] focused instead on the analysis and identification of color pigments using a digital camera, which served as a non-invasive, inexpensive, and portable tool for studying large surfaces. In their contribution, they propose a new supervised approach to camera characterization and color correction based on clustered data in which pigments are grouped based on their color or chemical properties. Fanfani et al. [4] focused on the restoration of historical photos, which offer valuable information and is an important source for art historians as they allow them to keep track of the changes that have occurred in a community and its 4.0/).



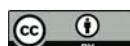
**Citation:** Bellavia, F.; Castellano, G.; Vessio, G. Editorial for Special Issue “Fine Art Pattern Extraction and Recognition”. *J. Imaging* **2021**, *7*, 195. <https://doi.org/10.3390/jimaging7100195>

Received: 26 September 2021

Accepted: 26 September 2021

Published: 29 September 2021

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

living environment over time. To this end, they present a fully automatic method for the digital restoration of historical stereo photos, which exploits the redundancy of the content in stereo pairs to detect and correct defects in the original images while improving contrast and illumination. Amura et al. [5] focused on graphic documentation, which refers to the systematic collection of information derived from diagnostic investigation as well as the restoration and monitoring processes. To facilitate and improve this investigation, and to drastically reduce manual interventions, they propose a semi-automatic methodology aimed at generating an objective and accurate graphic documentation to plan restoration, monitoring, and conservation interventions.

In their contribution, Banfi and Mandelli [6] exploited computer vision in combination with drone photogrammetry to develop a digital model of the Arco della Pace monument in Milan for augmented reality applications. The proposed method can improve interactivity and the sharing of information between users and digital heritage models, as well as the accessibility of details that would not be visible from the ground (especially the sculpture that crowns the top of the building).

Other contributions have addressed the problem of high-level semantic analysis and the interpretation of artworks. With the advent of digitized art collections, such an analysis has acquired increasing importance as a means of providing new information within digital art image repositories, supporting both enthusiasts and experts in finding and comparing artworks. To this end, computer vision techniques are good candidates to aid in the automatic categorization and retrieval of artworks. Following this line of research, Pincirol Vago et al. [7] experimentally compare several Class Activation Map techniques, which emphasize the areas of an image that contribute the most to the final classification performed by a convolutional neural network. This effort represents a step towards the creation of a computerized tool that is capable of highlighting variations in the positioning of iconographic elements, particularly for the detection of iconographic symbols in art images. Ghosh et al. [8] also focused on fine art classification, specifically proposing a method based on deep learning to classify geometric forms such as triangles and squares in mosaics. As a case study, a Roman mosaic is considered, which is digitally reconstructed by close-range photogrammetry based on standard photos. On the other hand, Tashu et al. [9] propose a multi-modal neural network based on sequential co-attention to classify artworks according to the emotions aroused in the observer. Emotion recognition in artworks is relevant as it can be used not only to group artworks, but also to provide recommendations that accentuate or balance a particular mood, or to find artworks of a specific style or genre that describe user-defined content in a user-defined affecting state. In her contribution, Cetinic [10] investigated the challenging problem of artwork captioning, which is the automatic generation of accurate and meaningful textual descriptions of artworks. To address this issue, she presents a captioning system developed by fine-tuning a transformer-based visual-language model. The results obtained suggest that it is possible to generate iconographically significant captions that capture not only the objects depicted, but also the historical and artistic context of an artwork. Abgaz et al. [11] also focused on the semantic enrichment of digitized cultural images and introduce a methodology for fully exploiting latent cultural information that is communicated visually by applying a combination of computer vision and semantic web technologies. A case study on food images is presented.

To support art historical research, Sindel et al. [12] did not focus on high-level concepts but on the distance measurement of the chain lines in historical prints, which constitute a sort of unique “fingerprint” of their paper structure. Since this process is typically manual, they propose an end-to-end trainable model based on a conditional generative adversarial network that performs line segmentation and parameterization in a multitask fashion.

We express our sincere gratitude to the authors for their contributions, to the reviewers for their efforts in reviewing the manuscripts, and to the editorial staff of the MDPI *Journal of Imaging* for their endless support in making this Special Issue possible. We hope it will benefit the scientific community and increase interest in this exciting area of research.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Del Bimbo, A.; Cucchiara, R.; Sclaroff, S.; Farinella, G.M.; Mei, T.; Bertini, M.; Escalante, H.J.; Vezzani, R. *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, 10–15 January 2021, Proceedings, Part III*; Springer Nature: Cham, Switzerland, 2021; Volume 12663.
2. Daffara, C.; Marini, E. A Portable Compact System for Laser Speckle Correlation Imaging of Artworks Using Projected Speckle Pattern. *J. Imaging* **2020**, *6*, 119. [[CrossRef](#)] [[PubMed](#)]
3. Trombini, M.; Ferraro, F.; Manfredi, E.; Petrillo, G.; Dellepiane, S. Camera Color Correction for Cultural Heritage Preservation Based on Clustered Data. *J. Imaging* **2021**, *7*, 115. [[CrossRef](#)]
4. Fanfani, M.; Colombo, C.; Bellavia, F. Restoration and Enhancement of Historical Stereo Photos. *J. Imaging* **2021**, *7*, 103. [[CrossRef](#)]
5. Amura, A.; Aldini, A.; Pagnotta, S.; Salerno, E.; Tonazzini, A.; Triolo, P. Analysis of Diagnostic Images of Artworks and Feature Extraction: Design of a Methodology. *J. Imaging* **2021**, *7*, 53. [[CrossRef](#)] [[PubMed](#)]
6. Banfi, F.; Mandelli, A. Computer Vision Meets Image Processing and UAS PhotoGrammetric Data Integration: From HBIM to the eXtended Reality Project of Arco della Pace in Milan and Its Decorative Complexity. *J. Imaging* **2021**, *7*, 118. [[CrossRef](#)]
7. Pincioli Vago, N.O.; Milani, F.; Fraternali, P.; da Silva Torres, R. Comparing CAM Algorithms for the Identification of Salient Image Features in Iconography Artwork Analysis. *J. Imaging* **2021**, *7*, 106. [[CrossRef](#)]
8. Ghosh, M.; Obaidullah, S.M.; Gherardini, F.; Zdimalova, M. Classification of Geometric Forms in Mosaics Using Deep Neural Network. *J. Imaging* **2021**, *7*, 149. [[CrossRef](#)] [[PubMed](#)]
9. Tashu, T.M.; Hajiyeva, S.; Horvath, T. Multimodal Emotion Recognition from Art Using Sequential Co-Attention. *J. Imaging* **2021**, *7*, 157. [[CrossRef](#)] [[PubMed](#)]
10. Cetinic, E. Towards Generating and Evaluating Iconographic Image Captions of Artworks. *J. Imaging* **2021**, *7*, 123. [[CrossRef](#)] [[PubMed](#)]
11. Abgaz, Y.; Rocha Souza, R.; Methuku, J.; Koch, G.; Dorn, A. A Methodology for Semantic Enrichment of Cultural Heritage Images Using Artificial Intelligence Technologies. *J. Imaging* **2021**, *7*, 121. [[CrossRef](#)] [[PubMed](#)]
12. Sindel, A.; Klinke, T.; Maier, A.; Christlein, V. ChainLineNet: Deep-Learning-Based Segmentation and Parameterization of Chain Lines in Historical Prints. *J. Imaging* **2021**, *7*, 120. [[CrossRef](#)]



Article

# A Portable Compact System for Laser Speckle Correlation Imaging of Artworks Using Projected Speckle Pattern

Claudia Daffara \* and Elisa Marini

Department of Computer Science, University of Verona, Strada le Grazie 15, 37134 Verona, Italy;  
elisa.marini.4@studenti.unipd.it

\* Correspondence: claudia.daffara@univr.it

Received: 5 August 2020; Accepted: 4 November 2020; Published: 6 November 2020



**Abstract:** Artworks have a layered structure subjected to alterations caused by various factors. The monitoring of defects at sub-millimeter scale may be performed by laser interferometric techniques. The aim of this work was to develop a compact system to perform laser speckle imaging *in situ* for effective mapping of subsurface defects in paintings. The device was designed to be versatile with the possibility of optimizing the performance by easy parameters adjustment. The system exploits a laser speckle pattern generated through an optical diffuser and projected onto the artworks and image correlation techniques for the analysis of the speckle intensity pattern. A protocol for the optimal measurement was suggested, based on calibration curves for tuning the mean speckle size in the acquired intensity pattern. The system was validated in the analysis of detachments in an ancient painting model using a short pulse thermal stimulus to induce a surface deformation field and standard decorrelation algorithms for speckle pattern matching. The device is equipped with a compact thermal camera for preventing any overheating effects during the phase of the stimulus. The developed system represents a valuable nondestructive tool for artwork diagnostics, allowing the monitoring of subsurface defects in paintings in out-of-laboratory environment.

**Keywords:** laser speckle imaging; speckle pattern; digital image correlation; nondestructive technique; artwork diagnostics; cultural heritage; portable system

---

## 1. Introduction

Artworks are subjected to structural alterations induced by various factors such as aging, microclimatic conditions and conservation treatments. In particular, ancient paintings present a complex layered structure that is susceptible to surface and subsurface decay, such as cracks, delaminations and detachments. The monitoring of such “defects” at small scale (sub-millimeter) is one of the objectives of nondestructive testing techniques applied to the conservation field [1,2].

Holographic interferometry [3] is a powerful technique that allows the detection of surface displacements with sub-micrometric accuracy. It is non-contact and non-invasive, highly sensitive and wide-field. Moreover, algorithms for processing image data are available, allowing fast and quantitative nondestructive analysis of structural modifications of the object in real-time [4]. The drawback of interferometry, generally speaking, is its sensitivity to external vibrations, which makes achieving the optimal measurement conditions without controlled laboratory settings difficult. This issue and the requirement of optics-skilled operators represent the major obstacle for a widespread use of such technique in the routine diagnostics of artworks. More flexible interferometry-based techniques are represented by the speckle-based methods [5], such as Electronic Speckle Pattern Interferometry (ESPI) and Speckle Pattern Photography (SPP) [6,7].

The effectiveness of holographic interferometry and speckle-based techniques for the analysis of artworks is well demonstrated, as reviewed in [8–13]. The applications include the structural evaluation of restoration processes (such as consolidation, cleaning or protective treatments), the detection of alterations induced by aging (such as cracks or subsurface defects) and the real-time monitoring of deformation due to microclimate variations [14–17]. As the nature of many artifacts makes their transport to a dedicated facility not possible, many research efforts are addressing the problem of in situ diagnostics with portable speckle-based techniques [14,18–26].

The speckle effect arises from the interaction of a coherent radiation with a random structure, as is the case of reflection from a rough diffusing surface or transmission across a diffusive medium [27], leading to an observed intensity pattern with a typical granular appearance, resulting from the multi-interference of the dephased waves at microscopic scale. In particular, any deformation of the surface micro-morphology turns into a modification of the corresponding anchored speckle pattern, down to displacements in the order of multiples of the radiation wavelength [6].

Subsurface defects in artworks, such as a lack of adhesion among the constitutive layers, can be detected by inducing an opportune thermal stress. In correspondence of the detachment, the heat dispersion rate slows down the return to equilibrium causing an irregular deformation field, observed at surface level.

Speckle metrology, in short, allows measuring the object deformations by acquiring and analyzing a sequence of speckle patterns. The ESPI technique is based on a two-beam configuration similar to holographic interferometry (as such, it is highly sensitive, up to sub-micron scale) and has found advantageous applications in the conservation field [8,13,15,28–30]. The SPP technique, conversely, is performed without the reference beam, by acquiring the speckle intensity pattern generated by the object beam alone. By correlation analyses of speckle images acquired from the object in different states, for example before and after a thermal *stimulus*, the deformation field can be obtained. SPP is especially sensitive to in-plane displacement components and local tilting [6], down to tenth micrometers. For a full characterization of the displacement field, speckle imaging and speckle shear interferometry can be integrated in a single device [19]. Despite the lower sensitivity with respect to ESPI, the SPP technique has the advantage of simpler setup, acquisition procedure and stability requirements. Speckle image decorrelation was demonstrated effective in detecting defects on wooden paintings, frescoes and mosaics, also in comparison with ESPI, as documented in early work [18,31,32].

While speckle diagnostics on interferometry basis with portable ESPI setup is being largely reported in the above-mentioned literature, it seems that less interest has been given to speckle imaging-based methods and SPP setup in the specific field of artwork diagnostics. Some recent works are concerned with laser speckle imaging for the dynamic analysis of material processing in restoration (drying and solvent actions [14,26]).

The aim of our work was the development of an effective, portable and compact system for performing speckle correlation imaging on artworks in situ. The system exploits an indirect speckle pattern projected onto the painting instead of the (more usual) speckle pattern generated by the surface. The paper is focused on the instrumentation and optimal performance setup and presents a validation of the measurement protocol on a layered painting sample to demonstrate that the developed system is well dimensioned and that it is effective in the analysis of subsurface defects.

## 2. Material and Methods

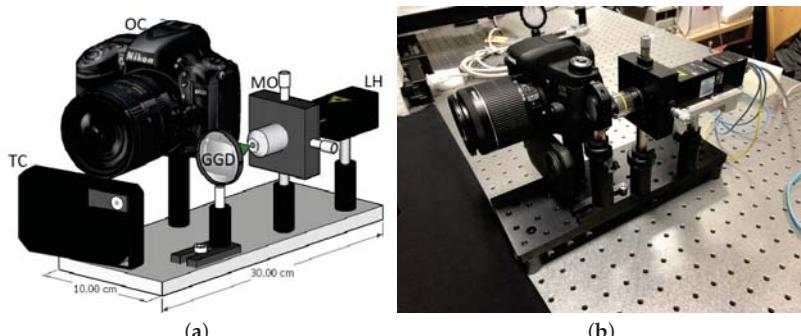
### 2.1. An Effective System for Speckle Pattern Correlation Imaging of Artworks

This paper presents an effective system for laser speckle imaging of artworks. The following key features have driven the design of the system:

- the hardware device should be compact and portable, with a versatile setup configuration, for performing measurements in out-of-lab environments;

- the overall measurement process should be completely noninvasive, with any effect potentially harmful to the artwork (namely, the heating) under control;
- the sensitivity performance should be versatile, tailorable to the specific diagnostics by easy parameters adjustment;
- a measurement protocol should be defined, and, ideally, the workflow should be as simple as possible also for operators with non-optics (interferometry) skills; and
- the hardware and the software should be commercially available and cost-effective, with the advantage of a system eventually at disposal of a wider conservation science community.

The designed system is shown in Figure 1. The setup is composed of the modules for generating, acquiring and processing the laser speckle pattern; the thermography module; and an external heating module. The modular design allows the fine adjustment of the components without affecting the alignment of non-involved parts. The laser source, thermal camera and photographic camera are fixed on a  $10 \times 30$  cm optical breadboard and can be easily accessed and moved independently of one another. The system is compact and portable for in situ diagnostics: the total weight is about 3 kg and the optical breadboard can be easily mounted on a stiff photographic tripod. Of course, the setup with the acquisition camera positioned in free-standing, in a separate tripod, is possible.



**Figure 1.** (a) Setup scheme. LH, laser head; MO, microscope objective; GGD, ground glass diffuser; OC, optical camera; TC, thermal camera. Approximate total weight: 3 kg. The supporting breadboard can be mounted on a stiff photographic tripod. (b) Picture of the device.

#### The laser speckle-generator module includes:

- a DPSS laser source, 532 nm, 125 W power-controlled (RGB lasersystem);
- a microscope objective, serving as beam expander for a wide field of illumination (up to  $1\text{ m}^2$ ) of the artwork; and
- an optical thin diffuser placed after the objective microscope and mounted on a stage to control the focused laser spot (as motivated later).

#### The speckle image acquisition module includes:

- a commercial high-resolution photographic camera able to image the speckle pattern (in this work we tested two CMOS cameras: Nikon D810 ( $7360 \times 4912$  pixel) coupled to a 50 mm lens with aperture  $f/9\text{--}11$ , controlled by the opensource software digiCamControl (Version 2.1.2) and Canon EOS 760D ( $6000 \times 4000$  pixel) coupled to a 18–55 mm zoom lens with aperture  $f/3.5\text{--}22$ , controlled by the smartphone app Camera Connect); and
- optical filters, with, optionally, a narrow-band filter for matching the laser wavelength and a linear polarizer for cleaning the speckle pattern.

#### The speckle image correlation module includes:

- software based on digital image correlation (DIC) for speckle pattern matching (in this work, we tested two standard methods, discussed below: a speckle decorrelation algorithm, used in cultural heritage applications [18], and a particle image velocimetry software, MatPIV (Version 1.7) [33], available as Matlab toolbox (Version R2014b))

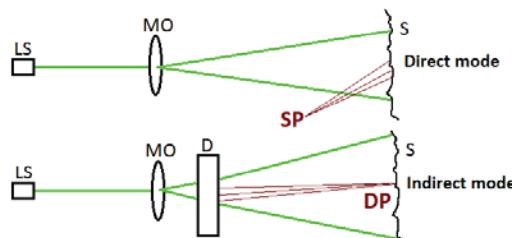
**The thermography module includes:**

- a bolometer-based thermal camera (FLIR C2,  $80 \times 60$  sensor in the infrared range  $7.5 \mu\text{m}$  to  $14 \mu\text{m}$  with thermal sensitivity  $<0.10^\circ\text{C}$ ), mounted next to the camera, for monitoring the effects of the thermal stimulus; and
- software (ResearchIR (Version 4.40.7.26)) for acquisition of radiometric thermal sequence in real-time.

**The external heating module includes:**

- a set of 750 W quartz tungsten infrared elements to apply a controlled step-heating pulse to the artwork, if the thermal stimulus is necessary; and
- a relay and a Theremino module to interface the lamps to the computer. It is important to shutter the lamps at the end of the stimulus, to avoid residual heating from the switch-off transient effect.

The system can be configured in two working modes: the indirect mode, in which the speckle pattern is generated by the diffuser in transmission geometry and projected on the artwork, and the direct mode, without the diffuser, in which the speckle pattern is generated in reflection geometry by the artwork surface (Figure 2).



**Figure 2.** Schematic setup in direct and indirect configuration. LS, laser source; MO, microscopic objective; D, diffuser; S, surface; DP, diffuser speckle pattern; SP, surface speckle pattern.

## 2.2. Setup Characterization Tests

Tailoring a speckle-based imaging application to the heritage field requires a thorough control of the setup as well as fine parameters tuning. The first laboratory tests were carried out to verify the suitability of the components in the main speckle-generator module: the laser source (effective output wavelength and wavelength variations with output power, beam divergence and polarization); the beam expander (suitability of microscopic objectives, with respect to transmission and expansion); the diffuser element (quality of the pattern, mainly with respect to transmission and diffusion level and intensity distribution, of different optical diffusers, from fine to coarse scattering).

The laser module was verified to produce a coherent, polarized and monochromatic beam. In the indirect configuration, a  $10\times$  microscopic objective (Olympus Plan Achromat (RMS10X)) coupled to the diffuser was able to provide a highly expanded beam (field of illumination of  $\sim 1 \text{ m}^2$  at  $1 \text{ m}$  distance), while higher magnification was needed in the direct configuration without the diffuser. Among the optical diffusers (Thorlabs N-BK7 Ground Glass; 120, 220, 600 and 1500 grit polishes), the finer grain was chosen for reducing the contribution from the zero-order beam, i.e., the component not diffracted away from the optical axis, preserving most of its energy and not properly diffused.

To address the problem of vibrations, the mirror-up mode of the camera must be enabled to avoid the blurring caused by the movement of the reflex mirror. Furthermore, the camera electronic front-curtain shutter could be employed instead of the mechanical one. In this regard, some tests were conducted, showing significantly reduced noise thanks to this simple expedient.

### 2.3. Motivation for the Thermography Module

Prolonged exposure to direct laser radiation during the measuring process could be harmful to delicate artifacts, in as far as it causes uncontrolled temperature rising. The thermal effect is mainly due to the directionality of the collimated laser beam and is reduced by the presence of the ground glass diffuser. Anyway, when applying the speckle technique to the analysis of subsurface defects, an external thermal solicitation is necessary for stressing the sample and inducing the displacement field. Following the request of the conservation scientists, to prevent any overheating effects in non-homogeneous artwork materials, we planned to acquire the artwork surface with a thermal camera, in continuous way, for the whole measuring session, thus allowing the temperature gradients to be monitored in real time. Beside monitoring the overheating, the thermal camera allows controlling the quality of the thermal stimulus during the excitation phase, i.e., that the infrared sources provide uniform and full-field irradiation and that the switch-off transient effect is effectively blocked (shuttering). The thermal load is quantified through a measurement of the raise of the surface temperature. Speckle methods were demonstrated on paintings using a weak thermal solicitation, with  $\Delta T$  not exceeding few degrees [10,29,31,32].

The FLIR C2 mounted in the system is a low-cost compact thermal camera, recently available in the market, with access to the radiometric data; the small-size of the sensor is not a limiting factor to our application as this camera also acquires a visible image superimposed to the thermal one.

### 2.4. Performance Analysis

Since the laser source is characterized by high spatial and temporal coherence, the pattern originated from (direct configuration) or projected onto (indirect configuration) the surface is very stable in time and thus provides a valid “fingerprint” of the object. On the one hand, the speckle pattern is generated from the surface asperities at microscopic scale; as such, it is intrinsic to the surface itself. On the other hand, the speckle pattern is generated by the diffuser, which is a stationary random medium, and then projected onto the object, where, again, multi-interference at wavelength scale occurs and a “speckled-speckle” [27] pattern is formed. In principle, both the direct and the indirect configuration can be used for probing the structure of the surface and its deformation in time at micrometric scale. However, the pattern generated by the surface in the direct configuration (without the diffuser) is expected to have finer granularity than the projected one, thus be more difficult to handle.

Coming to the application of artworks analysis *in situ*, some considerations are needed concerning the specific focus of the diagnostics, the most important one, if the investigation is aimed at detecting surface defects or subsurface defects. Even if a rigid classification is not possible when dealing with artworks, by surface defects, we mean those related to a degradation of the surface texture, e.g., abrasion and painting *craquelure*, while, by subsurface defects, we mean those related to material inhomogeneities in the deeper structure, e.g., detachments in the painting or in the plaster layer, voids and fillings. Mapping surface defects thus necessarily requires the speckle pattern to carry intrinsic information from the surface. The presence of subsurface defects, instead, may be detected in extrinsic way from the relative displacement field of the surface as response to a stimulus. In this case, both the direct surface speckle pattern and the indirect projected speckle pattern can be effectively used. The point is that, strictly speaking, when the projected speckle pattern of the diffuser is used, there is, always, also a secondary speckle pattern produced by the surface.

The main two aspects which we are interested in are concerned with the morphology of the pattern, namely the extent of the granularity, and with the contrast of the intensity pattern.

### 2.4.1. Sensitivity (Single Speckle Size)

The lower limit to the magnitude of the displacement that can be measured with the proposed setup is primarily determined by the mean speckle size, which in turn depends on the optical (lens aperture) and geometrical (distances) parameters of the setup. Following Goodman [34], the mean speckle size describes the extent of the spatial correlation of the speckle pattern and can be estimated as width of the autocorrelation function of the intensity at the observation plane. The pattern originated by the diffuser and projected on the artwork can be treated as “objective” speckles in free space (far field), in which the theoretical speckle size is given by [34]

$$s_{\text{obj}} \approx \frac{\lambda z}{D} \quad (1)$$

where  $\lambda$  is the laser wavelength,  $z$  is the object-to-viewing plane distance and  $D$  is the diameter of the region of the diffracting object (diffuser) illuminated by the laser.

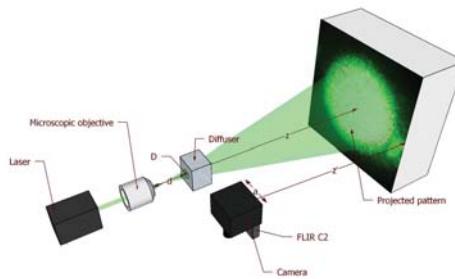
Concerning the secondary speckle pattern originated at the artwork surface plane, we have that the speckled-speckle field, after propagating in free-space, is imaged by the camera, therefore it turns into the so-called “subjective” speckle, whose final size is determined by the lens aperture [34]

$$s_{\text{subj}} \approx \frac{\lambda v}{a} \quad (2)$$

where  $v$  is now the lens-to-image plane distance and  $a$  is the aperture diameter of the lens of the imaging system.

If the system is used without the diffuser, in the direct configuration, a single pattern is formed from the rough surface of the artwork and then imaged by the camera, as subjective speckle, with the speckle size thus determined again by Equation (2).

Generally speaking, for the aim of artworks’ diagnostics, the speckle size should vary in the range 100  $\mu\text{m}$  to 700  $\mu\text{m}$  to ensure a good sensitivity. The setup is designed to allow a fine tuning of the projected speckle size by adjusting the working distance of the microscope objective that focuses the laser spot on the diffuser, thus controlling the parameter  $D$  (see the setup scheme in Figure 3). The projected pattern is then reflected by the surface and imaged by the camera, so the diffuser speckles’ size changes according to the magnification factor of the camera lens.

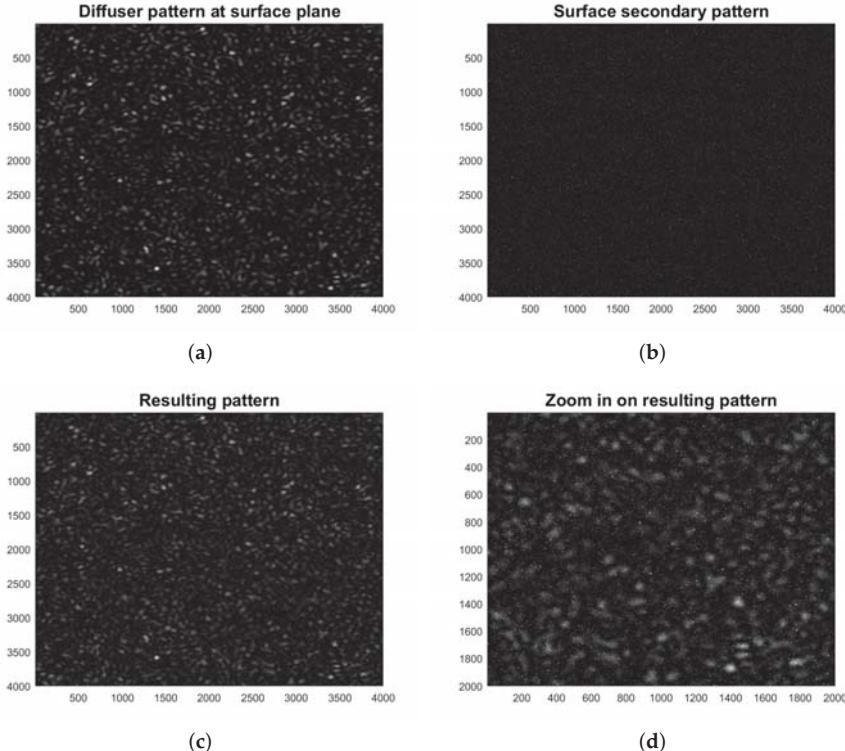


**Figure 3.** Setup scheme with the adjusting parameters: diffuser-to-object distance  $z$ ; laser spot  $D$  (through the microscope working distance  $d$ ); camera working distance  $z'$ ; and lens aperture  $a$ .

It should be remarked that the above formulas for the mean speckle size are obtained under the common assumption that the speckle intensity displays negative exponential statistics [27]. The actual resolution of the whole apparatus depends on the entire optical chain (mainly, effects from camera detector integration and pixel sampling, partial depolarization of the scattered light, and non-uniform reflective surface). For this reason, having complete control on the size of the generated speckles,

through an optimal parameters-tuning protocol, is fundamental for successive interpretation of the data. In this regard further studies are needed in order to get the full picture on how to originate patterns with specific, desired features by acting on the various parameters of the optical chain.

In Figure 4, we show the appearance of the projected and secondary patterns and their superposition resulting at the camera observation plane. The simulations of the objective diffuser pattern and the subjective surface pattern were performed following the authors of [35,36].



**Figure 4.** Speckle pattern simulations: (a) diffuser pattern at the surface plane; (b) surface secondary pattern; (c) superposition of the patterns as resulting after the reflection by the surface; and (d) zoom view.

#### 2.4.2. Optimization of Projected and Secondary Speckle Sizes

As explained above, the projected speckle pattern works as a kind of (random) structured light that is simply reflected by the artwork surface, which in turn generates its own speckle pattern. The resulting pattern observed through the camera of our setup appears as the superposition of two contributions: the indirect diffuser-generated pattern and the secondary surface-generated one (Figure 4c,d). Since the former is ultimately imaged after the reflection by the surface, the final ratio  $r$  between indirect speckles' size (taking into account the magnification of the lens in Equation (1)) and secondary speckles' size (Equation (2)) is independent of  $v$ :

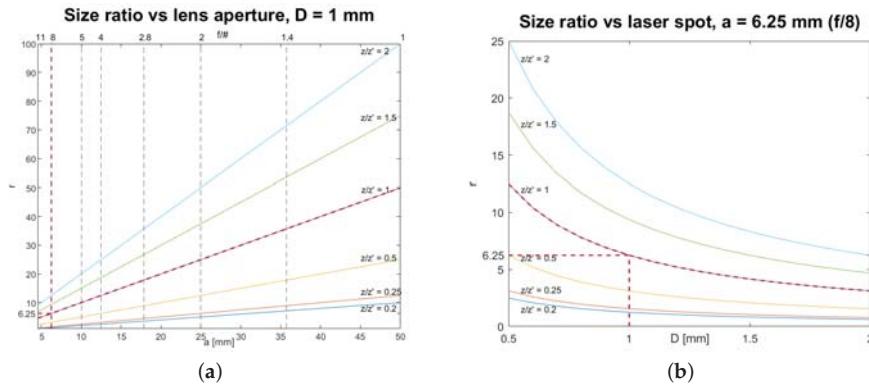
$$r \approx \frac{z}{z'} \frac{a}{D} \quad (3)$$

In this work, we focus on the detection of subsurface defects by exploiting the information carried by the projected pattern. To this aim, we have to adjust the parameters of the setup so as to

optimize the size of the diffuser speckles (the ones we are interested in) with respect to the surface ones. This way, we should minimize the information about the surface fine structure encoded in the acquired pattern while enhancing the signal related to the displacement field.

To aid the setting of the system, the behavior of the ratio  $r$  and its dependence on optical and distance parameters can be conveniently plotted and employed as indicative calibration curves. The following two practical situations are considered.

1. In the first case, as depicted in Figure 5a, we observe that, once the size of the projected speckles ( $D$  and  $z$  in Equation (1)) has been chosen and an adequate pixel size has been set through  $z'$  to detect them, we only need to move  $a$  to regulate the ratio  $r$ , which varies linearly. A greater value of  $D$  would cause the family of curves to shift downward, while the opposite effect would arise from a decrease in the fixed value of  $D$ .
2. The second case, as depicted in Figure 5b, shows how we can change the ratio between the two patterns by adjusting  $D$  and without changing the pixel size (since  $z'$  is fixed, the camera field of view (FOV) does not change). This way, we can increase or decrease (in the limit of the resolution allowed by the fixed pixel size) the size of the projected speckles with respect to that of the secondary surface speckles, acting only on the diffuser speckles' size. This is motivated by the fact that, once the field of illumination is chosen through the choice of  $z$  and the pixel size is determined through  $z'$ , the value of the relative speckle size of the two patterns can be varied arbitrary through the choice of  $D$  and it decreases hyperbolically as  $D$  increases. Increasing (respectively decreasing) the value of  $a$  would shift the family of curves upward (respectively, downward).



**Figure 5.** Calibration curves for the diffuser-over-surface speckles' size ratio  $r$  for various values of  $z/z'$ . (a) Fixing the laser spot  $D$  and moving the lens aperture  $a$ . On the bottom axis is the absolute lens aperture values and on the top axis is the f-stop values,  $f/\#$ , which can be directly chosen from the camera, given for the case of the 50mm lens. (b) Fixing the lens aperture  $a$  and moving the laser spot  $D$ . Red lines indicate the setup of the validation experiment ( $z = z'$ ) described in Section 3: for  $f = 50 \text{ mm}$ ,  $r = 6.25$  is achieved with f-number  $f/8$ .

## 2.5. Optimal Measurement Workflow

The above analysis suggests the following practical workflow for an optimal tuning of the setup parameters in the measurement session.

1. Identify the area to be inspected on the artwork surface. This will determine the field of illumination.

2. Set the diffuser-to-object distance  $z$  so as the diffuser projects the laser beam to uniformly cover the chosen area. According to Equation (1), the wider we make the field of illumination, by increasing  $z$ , the larger will result the projected speckles.
3. Set the laser spot  $D$  so that the diffuser-generated speckles have the desired size at the surface plane. This can be done by changing the distance  $d$  between objective and diffuser in the setup;  $D$  can be increased and tuned at will (up to diffuser diameter) to adjust the speckle size.
4. Set the camera-to-object distance  $z'$  to adjust the FOV and the pixel size. Decreasing  $z'$  will shrink the camera FOV increasing the resolution. In particular, the value of  $z'$  should allow to resolve details up to the order of the projected speckles' size. More specifically, according to the Nyquist criterion, a pixel size should be such that there are at least two pixels per speckle, being the speckles' size calculated on the basis of the needed diagnostic resolution.
5. The ratio of distances  $z/z'$  can be fixed as required by the out-of-lab condition, e.g., a museum environment, and the indirect to direct speckles' size ratio  $r$  can be controlled through the laser spot  $D$ , or through the lens aperture  $a$ , following the calibration curves, respectively, in Figure 5a,b.
6. In the case of environment with limited or restricted spaces, for example in presence of scaffolding, the compact system configuration with camera and laser mounted in the same optical breadboard can be employed, corresponding to  $z/z' \approx 1$ .

## 2.6. Speckle Image Correlation Analysis

As mentioned above, the data processing relies on two standard algorithms for pattern matching adapted to the SPP technique; the core of them is a local correlation principle. The first one, the speckle correlation (SC) algorithm [18], takes as input the intensity patterns acquired before and after the thermal stimulus and returns a correlation map where bright areas represent regions of high decorrelation, associated to anomalous in-plane displacements, while dark areas indicate regions where the pattern was almost unperturbed. The second one, MatPIV [33], is a particle image velocimetry tool, employed for the reconstruction of the surface average displacement field after the stimulus.

Briefly, the first algorithm works as follows: the speckle pattern acquired after the stimulus,  $I_{\text{mod}}(x, y)$ , is subtracted from the one acquired before,  $I_{\text{ref}}(x, y)$ , and then the difference is squared and averaged over an area containing many speckles, giving:

$$Q(x, y) = \langle [I_{\text{ref}}(x, y) - I_{\text{mod}}(x, y)]^2 \rangle \quad (4)$$

Under the assumptions of equal average intensities for the two patterns (due to stationarity),  $\langle I_{\text{ref}}(x, y) \rangle = \langle I_{\text{mod}}(x, y) \rangle = \langle I(x, y) \rangle$ , and of negative exponential distribution for the acquired intensity patterns (fully developed speckle assumption),

$$\rho_c(x, y) = \frac{Q(x, y)}{\langle I^2(x, y) \rangle} \quad (5)$$

is the complement-to-one of the (local) correlation coefficient  $\rho(x, y)$  of the intensity values measured at a point  $(x, y)$  of the two specklegrams, defined as the normalized (local) correlation function of the two patterns. The SC algorithm computes  $\rho_c(x, y)$  and plots it as a correlation map for the two specklegrams.

In practice, the calculation of the complement-to-one of the correlation coefficient was performed through a discrete convolution of the image matrix with a small matrix (kernel), which performs an average in the neighborhoods of image points, as suggested in similar works [18].

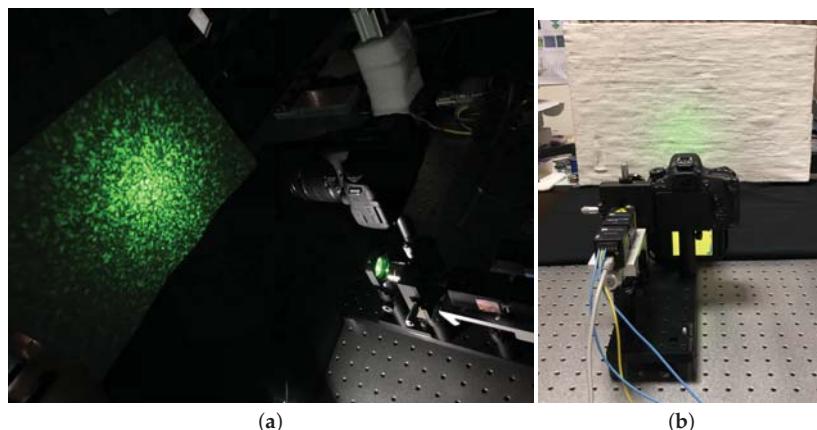
As regards the MatPIV software, both the acquired patterns are divided into corresponding sub-regions  $I_{\text{ref}}^{i,j}$  and  $I_{\text{mod}}^{i,j}$  and, for each pair, the components  $(u, v)$  of their relative in-plane displacement are estimated by minimizing the cross-correlation between  $I_{\text{ref}}^{i,j}$  and  $I_{\text{mod}}^{i,j}$ . After an

average displacement is associated to each pair of corresponding sub-regions, the final result comes from the juxtaposition of such local displacement fields. Local deformations of the surface are thus mapped as local irregularities in the direction or in the intensity of the average displacement field, in the same regions where the correlation between the two images was dropping before.

### 3. Results

The goal of this work was to propose an effective system for laser speckle pattern imaging (traditionally known as speckle photography) of artworks. Therefore, the developed prototype and the proposed optimal measurement protocol were validated in a typical context of artwork diagnostics by carrying out the experiment on a model of ancient painting with known hidden defects. The sample has the typical stratigraphy of Renaissance paintings and was prepared from the original receipts on materials and execution technique of *The Book of the Art* by Cennino Cennini [37], an Italian ancient treatise on paintings. A detailed description of the sample is given in Appendix A. The structural subsurface defects were modeled by inserting materials to break layer adhesion at various levels of the painting stratigraphy and in different positions, as shown in Figure A1.

The system was used in the compact configuration, with the modules mounted in the same breadboard, and in indirect mode, i.e., projecting the speckle pattern on the painting surface. The two photographic cameras with fixed-focus and zoom lens, as described above, were tested. A working distance of 50 cm with a laser power of 60 mW was set to have a field of illumination of  $\approx 40$  cm diameter with a good quality projected speckle pattern, well matched to the FOV of the Canon camera at resolution of 40  $\mu\text{m}$  (pixel size at object plane), assuring the optimal sampling of a 80  $\mu\text{m}$  minimum detail (speckle size), suitable for the detection of a typical defective region in panel paintings. The results obtained with the Canon camera are comparable to those obtained with the Nikon camera, where the focal length was fixed at 50 mm and we adopted the same parameters, namely, an aperture value of  $f/8$  and a working distance 50 cm, so as to have the same FOV with a higher resolution. A picture of the experiment setup is shown in Figure 6.



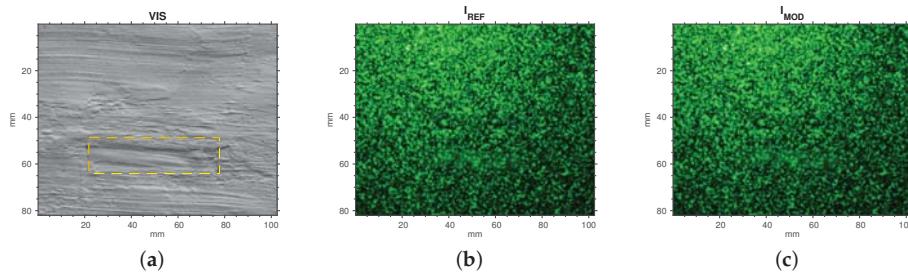
**Figure 6.** (a) Measurement setup with the system with camera, thermal camera and speckle module mounted in the same breadboard; and (b) back view with the lab lights on.

The thermal stimulus was applied to the painting using a single 750 W quartz tungsten infrared emitter. As discussed in Section 4.3, the duration of the thermal stress and the “response time” were critical parameters to optimize. Indeed, similar to any other active technique, i.e., based on the response of the system to an external solicitation (for example, infrared thermography), it is not possible to standardize such parameters in case of complex (and unknown) objects such as artworks. However,

by quantifying the thermal load in the raise of the surface temperature, measured by the thermal camera, the duration of the stimulus can be set and initially suggested, for example, by previous works [10,29,31,38]. Therefore, different experiments were performed varying the duration of the heat stimulus (5 s, 15 s and 30 s) up to a maximum  $\Delta T \sim 5^\circ\text{C}$  and acquiring the sequence of speckle patterns during the relaxation phase, after the lamp switch-off. In the case of the painting model, a short-pulse stimulus (5 s) was the most effective. This also satisfied the requirement of noninvasiveness.

In the design of DIC measurements, we followed the recommendations for the 2D-DIC given in “A Good Practices Guide for Digital Image Correlation” [39] by The International Digital Image Correlation Society. The object was positioned perpendicular to the optical axis; however, the surface of an ancient painting is not flat. The imaging system was used without any automatic adjusting, as auto-focus or apertures. As the employed DSLR cameras have an anti-alias filter, the images were not filtered. The intensity of the recorded speckle pattern was optimized by tuning the laser power, while keeping the gain of the camera low to minimize camera noise, as recommended.

The analysis was performed on a region of the painting with a chosen defect (#5 in Figure A1). Figure 7 depicts the Region Of Interest (ROI) with a pair of speckle intensity pattern processed in the DIC analysis, the reference pattern  $I_{\text{ref}}$  (before the thermal stimulus) and the modified pattern  $I_{\text{mod}}$  (after the thermal stimulus). From the calibration curve (Figure 5), the lens setting provides an indicative ratio of the indirect to direct speckles of  $r = 6.25$ .



**Figure 7.** (a) ROI analyzed by the algorithms and defect position. Example of processed specklegrams: (b) before stimulus, used as reference pattern; and (c) after (5 s) pulse stimulus, modified pattern. Laser: 60 mW. Camera lens: 50 mm; exposure: 1/8 s; aperture: f/8; ISO: 100. Laser spot (D): 1 mm; diffuser-to-surface distance (z): 50 cm; camera-to-surface distance (z'): 50 cm.

### 3.1. Results with the SC Algorithm

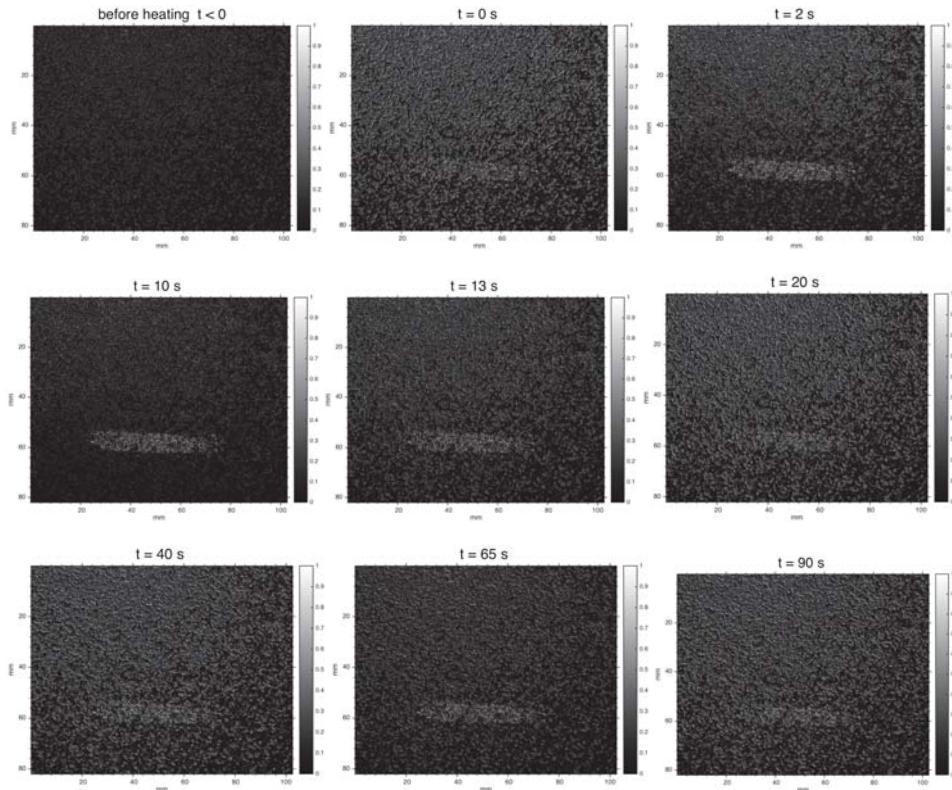
Figure 8 reports the correlation map computed by the SC algorithm for some representative frames. There is also a video (100 s) available showing the behavior of the correlation map over time (Supplementary Material, Video S1: laser speckle decorrelation maps sequence). The sequence was obtained by processing, at significant step, the pairs of specklegram  $I_{\text{mod}}(t)$ ,  $I_{\text{ref}}$ . Even if a continuous acquisition of the speckle activity is performed, one-beam speckle-intensity correlation methods require the off-line processing phase.

As one can see, a discontinuity in the gray level in the SC algorithm output of the normalized correlation coefficients (Equation (5)) well locates the defective region and its extension, where the painting surface undergoes an anomalous displacement. As expected, correlation computed on a pair of specklegrams of the sample in equilibrium before the thermal pulse does not reveal the defect. The bright decorrelated pixels that also appear in other regions, supposed to be non-defective, may be attributed to proper material inhomogeneities due to the handmade nature of the sample, as well as to external noise. As mentioned above, the effectiveness of speckle photography is based on the fact that the bulk defect, after thermal loading, determines a local irregular deformation of the surface and thus of the laser speckle activity. We see that in the case of pictorial detachments, which are

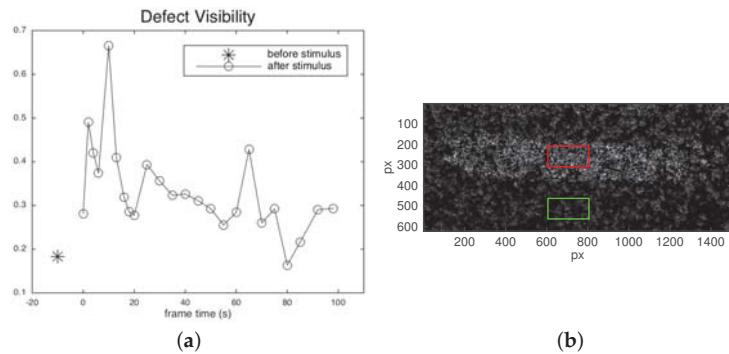
positioned immediately beneath the surface, the defect is early detected at the beginning of the relaxing phase. The maximum visibility is observed at  $t = 10$  s, persisting across the 100 s sequence with a contrast that varies as the deformation of the defective region behaves different from the entire regular surface. In the correlation map computed by the SC algorithm, this can be quantified in the following visibility parameter

$$\nu = \frac{\langle \rho_c \rangle_{\text{def}} - \langle \rho_c \rangle_{\text{ref}}}{\langle \rho_c \rangle_{\text{def}} + \langle \rho_c \rangle_{\text{ref}}} \quad (6)$$

that estimates the contrast in the correlation coefficient averaged in defective and reference (sound) region (Figure 9).



**Figure 8.** DIC results by SC algorithm. Laser speckle decorrelation map at different times:  $t = -10$  s, before thermal pulse stimulus (5 s); lamps switch-off at  $t = 0$ ;  $t > 0$ , cooling phase.

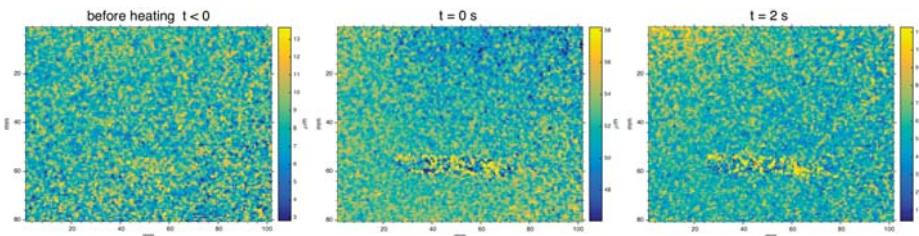


**Figure 9.** (a) Defect visibility at different frames, computed as Michelson contrast between the correlation coefficient averaged on a defective and on a sound ROI (Equation (6)).  $t = -10\text{ s}$ , before 5 s stimulus;  $t > 0$ , relaxing phase after stimulus. (b) Zoom of the correlation map with the selected ROIs ( $100 \times 200$  points) on defective (red) and sound (green) region.

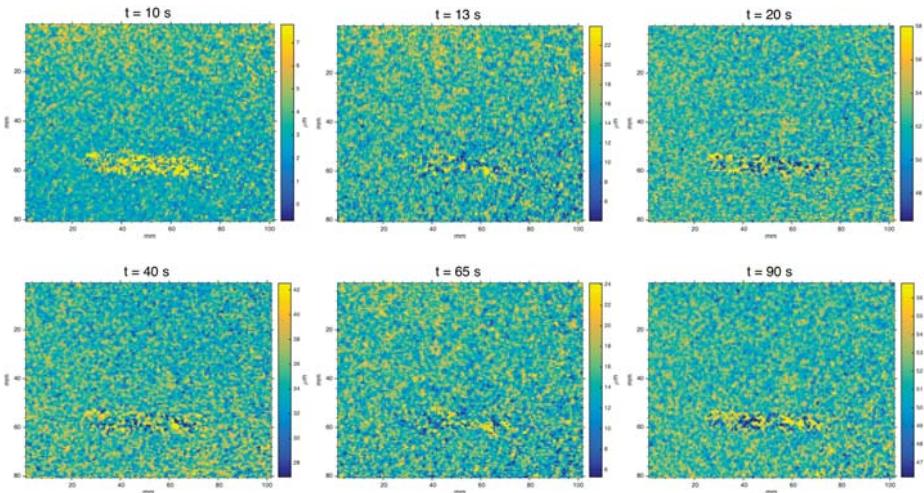
### 3.2. Results with the MatPIV Software

Figure 10 reports the displacement maps computed by the software MatPIV for some representative frames. There is also available a video (100 s) showing the behavior of the displacement field over time (Supplementary Material, Video S2: spatial displacement maps sequence)

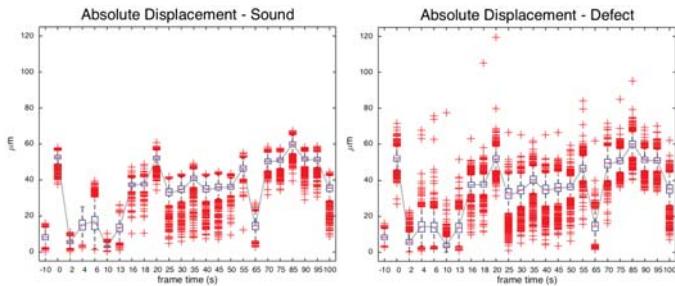
MatPIV algorithm is successful too in detecting the defect, through the displacement map at sub-millimetric scale. The defect is early detected after the pulse stimulus as local anomalous behavior of the displacement field, and the visibility persists across the 100 s sequence. As expected, MatPIV computed on pair of specklegrams of the sample in equilibrium before the thermal pulse does not reveal the defect. Noise-floor analysis in static images ( $t < 0$ ) gives a mean  $\mu \sim 8 \mu\text{m}$  and a variance error  $\sigma \sim 2 \mu\text{m}$ , mainly due to external vibration and camera noise. After the stimulus, bias errors are induced by heat waves, out-of-plane motion and vibration of the painting (large sample positioned in a vertical position). Anyway, it is interesting to examine the displacement distribution in the MatPIV maps. Figure 11 reports the boxplot of the absolute displacement over time, showing how the defective and reference sound regions exhibit the same mean displacement but a different dispersion in the relaxing phase. The coefficient of variation (relative standard deviation  $\sigma/\mu$ ) can be taken as indicator of the visibility of defect (Figure 12). The maximum visibility is observed at  $t = 10\text{ s}$ , similar to for the SC results, corresponding to small displacements but with a strong “decoupled” behavior of the local distribution in the defective region with respect to the regular surface.



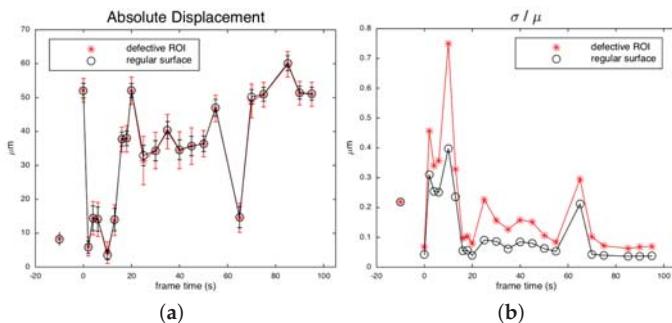
**Figure 10. Cont.**



**Figure 10.** Absolute displacement map computed by MatPIV at different times:  $t = -10$  s, before thermal pulse stimulus (5 s); lamps switch-off at  $t = 0$ ;  $t > 0$ , cooling phase. The LUT is set to  $\mu \pm 3\sigma$  for best defect visualization. Noise-floor analysis in static images gives a mean  $\mu \sim 8 \mu\text{m}$  and a variance error  $\sigma \sim 2 \mu\text{m}$ .



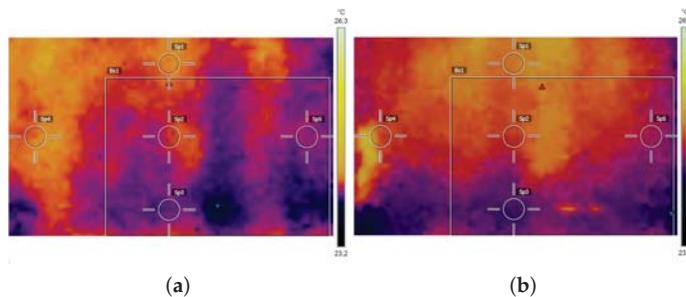
**Figure 11.** Boxplot to give an indication of the distribution of absolute displacements in defective and regular sound regions over time:  $t = -10$  s, before 5 s stimulus;  $t > 0$ , relaxing phase after stimulus.



**Figure 12.** (a) Mean absolute displacement for defective and sound ROIs over time; and (b) coefficient of variation of the displacement distribution for the defective ROIs and for the regular surface over time.

### 3.3. Application of the Thermographic Module

With a thermal stimulus of 5 s and a lamp-to-surface distance of 40 cm, we observed a temperature increment of  $\sim 1^{\circ}\text{C}$  right after the thermal solicitation (Figure 13). After 30 s waiting, the mean temperature was almost completely back to initial equilibrium. The mean temperature showed a maximum oscillation of  $1^{\circ}\text{C}$  during the whole measurement process. When the laser alone was turned on, we measured an increase in temperature  $<1^{\circ}\text{C}$ . Thermal emissivity was set to  $\varepsilon = 0.90$  for the involved material (gypsum). Being artworks surface characterized by inhomogeneous materials, roughness and decay degree, an assessment of the emissivity map for the calculation of the temperature field could be difficult. However, for many background materials, the tabulated values can be used.



**Figure 13.** Thermal data: (a) ( $\text{Sp1} = 24.2^{\circ}\text{C}$ ;  $\text{Sp2} = 24.1^{\circ}\text{C}$ ;  $\text{Sp3} = 23.9^{\circ}\text{C}$ ;  $\text{Sp4} = 24.1^{\circ}\text{C}$ ;  $\text{Sp5} = 24.0^{\circ}\text{C}$ ).  $\text{Bx1: max} = 24.2^{\circ}\text{C}$ ;  $\text{min} = 23.8^{\circ}\text{C}$ ;  $\text{average} = 24.0^{\circ}\text{C}$ ) before and (b) ( $\text{Sp1} = 24.9^{\circ}\text{C}$ ;  $\text{Sp2} = 24.7^{\circ}\text{C}$ ;  $\text{Sp3} = 24.3^{\circ}\text{C}$ ;  $\text{Sp4} = 24.7^{\circ}\text{C}$ ;  $\text{Sp5} = 24.6^{\circ}\text{C}$ ).  $\text{Bx1: max} = 24.9^{\circ}\text{C}$ ;  $\text{min} = 24.1^{\circ}\text{C}$ ;  $\text{average} = 24.6^{\circ}\text{C}$ ) after the thermal stimulus of 5 s. The box indicates the investigated ROI.

## 4. Discussion

### 4.1. Discussion on Contrast Sensitivity

In the performance analysis, we focused on the mean speckle size of the speckle pattern. The other issue affecting the overall system sensitivity is the contrast sensitivity, which is primarily determined by the contrast of the speckle pattern, commonly defined as  $C = \sigma_1/\mu_1$ , the ratio of the standard deviation to the mean of the intensity distribution. It is well known [27] that a polarized and fully developed speckle field displays negative exponential intensity distribution with speckle contrast  $C = 1$ .

The assumption of underlying negative exponential statistics for the acquired pattern is not always verified in practice, due to many factors such as the presence of the two scattering media (diffuser and object surface), the depolarization of the beam due to diffuse reflection at artwork surface, the integration of the camera sensor and the non-uniform reflectance typical of an artwork surface. It has been shown that [27]: the speckled-speckle pattern displays a conditional exponentially-driven intensity distribution with contrast  $C = \sqrt{3}$ ; depolarization causes a degradation of the contrast due to the sum of independent (non-coherent) patterns; and the sum in intensity of patterns displays a Gamma density function, with contrast that decreases as the time of integration increases. Together with the speckle pattern morphology, the measured speckle contrast is thus affected by the actual statistics. Moreover, it is degraded by the camera acquisition process; anyway, it has been shown that, if care is taken, namely by sampling the speckle size above the Nyquist criterium, the contrast of the acquired speckle image approaches the theoretical limit and is maximized [40].

We addressed some of these issues with the aim of reducing their impact on the ideal negative exponential intensity distribution and to verify how the unavoidable deviation from the latter affects the overall performance of the technique. In doing this, the ratio between the diffuser speckles

(expected to be fully developed) and the surface ones was set well over unity, so as to minimize the contribution of these latter to the acquired pattern. We placed a linear polarizing filter before the camera, after having verified that the diffuser does not depolarize the laser beam. The drawback, however, was the drop in intensity. As regards the integration process of the camera, numerical simulation tests have shown us that the statistics of the final pattern, and in particular its deviation from the exponential one, does not affect the performance of the software adopted for the DIC.

A study on the statistics of the intensity distribution and on the contrast of the speckle pattern in SPP application to artworks is planned as future development of this research.

#### 4.2. Discussion on Non-Uniform Reflectance

When dealing with artworks, a quite complex problem is the non-uniform reflectance of the polychromatic surface, as it affects the measured contrast locally. Figure 14 displays a speckle pattern projected on an ancient oil panel painting (anonymous, 16th century), showing the darker regions to absorb more incident radiation so that the speckle pattern is less or not detectable in their surrounding. Since the pattern is the unique information carrier, this could lead to a loss in information about surface modifications in the involved areas. To some extent, this issue can be addressed by controlling the external illumination and the power of the laser, but it remains a problem in the case of polychromatic and delicate artwork materials.

The study of the effect of non-uniform reflectance surface in SPP application to artworks is planned as future research direction.



**Figure 14.** Non-uniform reflectance effect in a 16th century oil painting (anonymous, private collection): (a) visible image; and (b) speckle intensity pattern.

#### 4.3. Discussion on Thermal Stimulus and Response Time

The duration of the thermal pulse, and hence its intensity, is a critical parameter. In order for a sub-surface defect to manifest as speckle activity, the sample must undergo a physical stress strong enough to make evident, on the surface, the inhomogeneities due to the presence of bulk damages. The deeper is the defect, the more intense the solicitation must be. At the same time, a strict noninvasiveness requirement limits the thermal gradient on the surface to few degrees. The experiments pointed out that the general deformation to which the entire surface undergoes can be a detrimental factor, whenever it reaches the same order of magnitude as the defective area displacement. The ideal solicitation should maximize the deformation of the detachment, leaving at the same time the surrounding surface relatively unperturbed.

Since what is detected by SPP is a deformation of the surface, the response time is also a critical parameter. After the lamp switch-off, the energy stored in the surface must propagate to the inner defect. The resulting deformation intensity and time duration depend on the thermal capacity and expansion coefficient of the materials encountered in the stratigraphy. A detachment is a resistive defect that requires extra time to return to its equilibrium, and its deformed status, if recorded in this phase, emerges from the surrounding regions.

For the painting-like support, a  $\Delta T \sim 1^\circ\text{C}$  (5 s pulse) was able to induce the visibility of sub-surface macroscopic detaches. Longer stimuli (15 s and 30 s) did not improve the detectability; the most intense deformation field induced on the whole surface causes the local displacement of the defective area to be “hidden” by that of the regular surface. Unless the defect is very deep in the bulk, a short stimulus inducing a  $2^\circ\text{C}$  to  $3^\circ\text{C}$  gradient on the surface is suggested. Regarding the materials’ response, the experiments suggested an interval time in the order of 10 s for the superficial defects and in the order of 100 s for the deeper ones, in agreement with the literature. In the case of different support, i.e., from wooden to mural paintings, the conduction coefficients vary, and consequently the optimal acquisition times.

#### 4.4. Discussion on Quantitative DIC

As mentioned above, the DIC measurements were designed following the good practices guide [39]. However, in speckle correlation applied to analysis of hidden sub-surface defects in artworks, some specific considerations must be made. The final objective is not the calculation of the derived field quantities, e.g., the displacement, but the localization of the defective area, which is revealed from an anomalous local behavior of the surface speckle activity with respect to the background. Moreover, the position of the sub-surface defect, as well as its nature, is mostly unknown; it is not possible to determine the expected surface deformation on a local ROI *a priori*, and the analysis should ideally be full-field. As a consequence, we have to face a trade-off between a large FOV and spatial resolution. The influence of optical system resolution on DIC uncertainties is discussed in [41]. DIC computation can be affected by many factors, related not only to the acquisition system and the laser speckle pattern but also to end-user decisions, such as the selection of the subset size to track the displacements [42]. One critical issue is that, for typical 2D-DIC applications, it is assumed that the sample remains planar at constant stand-off distance. Here, instead, the painting surface is subjected also to out-of-plane motion induced by the heat stimulus. In the case of artworks, it is not possible to estimate the out-of-plane deformation to compensate the in-plane measurements, as recommended by the good practice. Moreover, such out-of-plane motion, traced as fictitious in-plane absolute displacements, conveniently contributes to the irregular local behavior of the defective regions. Regarding the optics, the use of a bi-telecentric lens or of a long focal length is suggested to compensate the effect, allowing to solve the problem of magnification shifting and distortion.

## 5. Conclusions

A portable and very compact system for laser speckle imaging of artworks *in situ* is presented. The device was designed to be versatile, tailored to the needs of the art diagnostics field, with the possibility of optimizing the sensitivity performance by easy parameters adjustment. The system can operate in the indirect mode, in which the speckle pattern is generated through an optical diffuser and projected onto the artwork, and in the direct mode, in which the speckle pattern is generated by the artwork surface. The optimization of the optical setup through a tuneable speckle size (direct-surface and indirect-diffuser) was designed after a theoretical analysis of the performance based on statistical optics. A protocol for the optimal measurement was suggested, based on calibration curves for obtaining the desired mean speckle size in the acquired intensity pattern.

The system was validated in the analysis of subsurface defects in a model of ancient painting, using a short pulse thermal stimulus to induce a surface deformation field and the image correlation technique for the analysis of the sequence of speckle intensity patterns. To demonstrate that the developed system was well dimensioned and effective, the DIC was performed using two standard methods: the Speckle Correlation (SC) and the Particle Image Velocimetry (PIV) algorithms. The thermal loading induces irregular surface and sub-surface micro-motion, in-plane and out-of-plane, which is not traced quantitatively, but, if a proper (limited) thermal stress is used, a proper speckle size is used and a proper interrogation subset is used in DIC, the SPP technique is effective in differentiating the defective region as irregular local behavior with respect to the background. Even if a quantitative

2D-DIC was not the final objective (and not possible with complex multi-layered artworks), an analysis of the visibility of the defect was made in the correlation image sequence by SC and in the PIV displacements maps.

Following the requirement of noninvasiveness, a compact thermal camera was mounted on the system for monitoring the temperature of the artwork. The thermal camera allowed to face the critical issue of the optimization of the thermal stimulus: On the one hand, it allowed a quantification of the load intensity through a measurement of the surface temperature and an initial setting of the duration of the pulse (for example, following similar works in literature). On the other hand, it allowed the thermal solicitation to be maintained within the safe range of conservation standards. A surface temperature gradient  $\sim 1$  °C induced by a short thermal pulse of 5 s was optimal for the detection of sub-surface detachments in the painting-like model, in accordance with the noninvasiveness requirement.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2313-433X/6/11/119/> S1, Video S1: laser speckle decorrelation maps sequence; Video S2: spatial displacement maps sequence.

**Author Contributions:** Conceptualization, C.D.; methodology, C.D.; investigation, C.D. and E.M.; formal analysis, C.D. and E.M.; data curation, E.M.; visualization, C.D. and E.M.; writing—original draft, C.D.; writing—review and editing, C.D.; resources, C.D.; and supervision, C.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Acknowledgments:** The painting sample used for experiments was fabricated by Luca Perlini.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

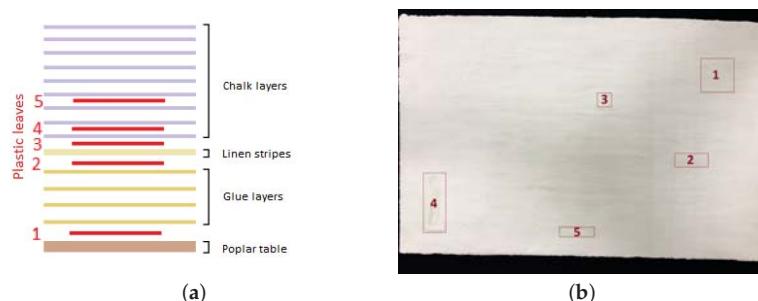
ESPI	Electronic Speckle Pattern Interferometry
SPP	Speckle Pattern Photography
DPSS	Diode-Pumped Solid-State
CMOS	Complementary Metal-Oxide Semiconductor
DIC	Digital Image Correlation
FOV	Field Of View
SC	Speckle Correlation
ROI	Region Of Interest

## Appendix A. The Painting Sample

The model of the Renaissance painting used in the experimental validation was fabricated following the traditional receipts on constitutive materials and execution technique contained in the ancient treatise *The Book of the Art* by Cennino Cennini [37]. The structural subsurface defects were obtained, as usual [10], by inserting materials with different thermal response.

We reproduced the process called *imprimitura*: a multi-layer poplar panel of  $40 \times 60$  cm, preliminarily treated with four hands of solution of rabbit-skin glue with different dilution degrees, was covered with a number of rough linen stripes after they were immersed in the glue solution as well. After a drying time of two days, the table was smoothed with fine grain sandpaper, then some Bologna chalk powder was mixed to the rabbit-skin glue and the resulting mixture was warmed-up in a bain-marie and laid on the panel. After a drying time of three days, the surface of the table was smoothed again and eight layers of chalk and glue mixture were laid on it, each just before the complete drying of the previous one. After two final days of drying, the surface was smoothed one last time, obtaining the final sample, whose stratified structure is shown in Figure A1.

During the process, five thin plastic leaves were inserted at various levels of the stratigraphy and at different positions so as to simulate detachments of known dimensions, depth, and positions (Figure A1).



**Figure A1.** The fabricated painting model ( $40 \times 60$  cm): (a) scheme of the layered structure with the defects; and (b) picture with the position of the defective regions (numbered in red boxes).

## References

- Alfeld, M.; Broekaert, J.A. Mobile depth profiling and sub-surface imaging techniques for historical paintings—A review *Spectrochim. Acta B* **2013**, *88*, 211–230. [[CrossRef](#)]
- Borg, B.; Dunn, M.; Ang, A.; Villis, C. The application of state-of-the-art technologies to support artwork conservation: Literature review. *J. Cult. Herit.* **2020**, *44*, 239–259. [[CrossRef](#)]
- Vest, C. *Holographic Interferometry*; Wiley & Sons: New York, NY, USA, 1979.
- Shimobaba, T.; Ito, T. *Computer Holography: Acceleration Algorithms and Hardware Implementations*; CRC Press: Boca Raton, FL, USA, 2019.
- Rastogi, P.K. (Ed.) *Digital Speckle Pattern Interferometry and Related Techniques*; John Wiley & Sons Inc.: New York, NY, USA, 2000; p. 384.
- Jones, R.; Wykes, C. *Holographic and Speckle Interferometry*; Cambridge University Press: Cambridge, UK, 1989. [[CrossRef](#)]
- Cloud, G. *Optical Methods of Engineering Analysis*; Cambridge University Press: Cambridge, UK, 1995. [[CrossRef](#)]
- Paoletti, D.; Spagnolo Schirripa, G. Interferometric Methods for Artwork Diagnostics. *Prog. Opt.* **1996**, *35*, 197–255.
- Tornari, V.; Bonarou, A.; Zafiropulos, V.; Fotakis, C.; Smyrnakis, N.; Stassinopoulos, S. Structural evaluation of restoration processes with holographic diagnostic inspection. *J. Cult. Herit.* **2003**, *4*, 347–354. [[CrossRef](#)]
- Ambrosini, D.; Paoletti, D. Holographic and speckle methods for the analysis of panel paintings. Developments since the early 1970s. *Stud. Conserv.* **2004**, *49*, 38–48. [[CrossRef](#)]
- Dulieu-Barton, J.M.; Dokos, L.; Eastop, D.; Lennard, F.; Chambers, A.R.; Sahin, M. Deformation and strain measurement techniques for the inspection of damage in works of art. *Stud. Conserv.* **2005**, *50*, 63–73. [[CrossRef](#)]
- Hinsch, K.D. Laser speckle metrology—a tool serving the conservation of cultural heritage. In *Oscillation, Waves, and Interaction*; Kurz, T., Parlitz, U., Kaatze, U., Eds.; Universitätsverlag Göttingen: Göttingen, Germany, 2007; pp. 259–278.
- Tornari, V. Laser interference-based techniques and applications in structural inspection of works of art. *Anal. Bioanal. Chem.* **2007**, *387*, 761–780. [[CrossRef](#)] [[PubMed](#)]
- Perez, A.; Gonzalez-Pena, R.; Braga, R., Jr.; Perles, A.; Perez-Marin, E.; Garcia-Diego, F. A Portable Dynamic Laser Speckle System for Sensing Long-Term Changes Caused by Treatments in Painting Conservation. *Sensors* **2018**, *18*, 190. [[CrossRef](#)]
- Hinsch, K.D.; Zehnder, K.; Joost, H.; Gölker, G. Monitoring detaching murals in the Convent of Müstair (Switzerland) by optical metrology. *J. Cult. Herit.* **2009**, *10*, 94–105. [[CrossRef](#)]
- Schirripa Spagnolo, G.; Ambrosini, D.; Guattari, G. Electro-optic holography system and digital image processing for in situ analysis of microclimate variation on artworks. *J. Opt.* **1997**, *28*, 99–106. [[CrossRef](#)]
- Tornari, V.; Bernikola, E.; Nevin, A.; Kouloumpis, E.; Doulgeridis, M.; Fotakis, C. Fully-Non-Contact Masking-Based Holography Inspection on Dimensionally Responsive Artwork Materials. *Sensors* **2008**, *8*, 8401–8422. [[CrossRef](#)]

18. Schirripa Spagnolo, G.; Ambrosini, D.; Paoletti, D. An NDT electro-optic system for mosaics investigations. *J. Cult. Herit.* **2003**, *4*, 369–376. [[CrossRef](#)]
19. Groves, R.M.; Fu, S.; James, S.W.; Tatam, R.P. Single-axis combined shearography and digital speckle photography instrument for full surface strain characterization. *Opt. Eng.* **2005**, *44*, 025602. [[CrossRef](#)]
20. Castellini, P.; Abaskin, V.; Achimova, E. Portable electronic speckle interferometry device for the damages measurements in veneered wood artworks. *J. Cult. Herit.* **2008**, *9*, 225–233. [[CrossRef](#)]
21. Lasyk, L.; Lukomski, M.; Bratasz, L. Portable electronic speckle interferometry device for the damages measurements in veneered wood artworks. *Opt. Appl.* **2011**, *41*, 688–700.
22. Boaglio, E.; Lamas, J.; López, A.J.; Ramil, A.; Pereira, L.; Prieto, B.; Silva, B. Development of a portable ESPI system for the analysis in situ of mural paintings. In *Speckle 2012: V International Conference on Speckle Metrology*; Doval, A.F., Trillo, C., Eds.; International Society for Optics and Photonics, SPIE: Bellingham, WA, USA, 2012; Volume 8413, pp. 243–248. [[CrossRef](#)]
23. Tornari, V. On development of portable Digital Holographic Speckle Pattern Interferometry system for remote-access monitoring and documentation in art conservation. *Strain* **2018**, *55*, e12288. [[CrossRef](#)]
24. Memmolo, P.; Arena, G.; Fatigati, G.; Grilli, M.; Paturzo, M.; Pezzati, L.; Ferraro, P. Automatic frames extraction and visualization from noisy fringe sequences for data recovering in a portable digital speckle pattern interferometer for NDI. *J. Disp. Technol.* **2015**, *11*, 417–422. [[CrossRef](#)]
25. Perlini, L.; Ambrosini, D.; Daffara, C. A versatile system for in-situ speckle and thermography-based diagnostics of artifacts. *IOP Conf. Ser. Mater. Sci. Eng.* **2018**, *364*, 012063. [[CrossRef](#)]
26. Baij, L.; Buijs, J.; Hermans, J.J.; Raven, L.; Iedema, P.D.; Keune, K.; Sprakel, J. Quantifying solvent action in oil paint using portable laser speckle imaging. *Sci. Rep.* **2020**, *10*, 10574. [[CrossRef](#)]
27. Goodman, J. *Speckle Phenomena in Optics: Theory and Applications*; Roberts & Company: Englewood, CO, USA, 2007.
28. Lucia, A.C.; Zanetta, P.M.; Facchini, M. Electronic speckle pattern interferometry applied to the study and conservation of paintings. *Opt. Laser Eng.* **1997**, *26*, 221–233. [[CrossRef](#)]
29. Albrecht, D.; Franchi, M.; Lucia, A.C.; Zanetta, P.M.; Aldrovandi, A.; Cianfanelli, T.; Riitano, P.; Sartiani, O.; Emmony, D.C. Diagnostic of the conservation state of antique Italian paintings on panel carried out at the Laboratorio di Restauro dell’Opificio delle Pietre Dure in Florence, Italy with ESPI-based portable instrumentation. *J. Cult. Herit.* **2000**, *1*, S331–S335. [[CrossRef](#)]
30. Krzemień, L.; Lukomski, M.; Kijowska, A.; Mierzejewska, B. Combining digital speckle pattern interferometry with shearography in a new instrument to characterize surface delamination in museum artefacts. *J. Cult. Herit.* **2015**, *16*, 544–550. [[CrossRef](#)]
31. Schirripa Spagnolo, G.; Ambrosini, D.; Paoletti, D. Image decorrelation for in situ diagnostics of wooden artifacts. *Appl. Opt.* **1997**, *36*, 8358–8362. [[CrossRef](#)]
32. Schirripa Spagnolo, G.; Paoletti, D.; Ambrosini, D.; Guattari, G. Electro-optic correlation for in situ diagnostics in mural frescoes. *Pure Appl. Opt.* **1997**, *6*, 557. [[CrossRef](#)]
33. Sveen, J.K. *An Introduction to MatPIV v. 1.6.1*; Department of Mathematics, University of Oslo: Oslo, Norway, 2004.
34. Goodman, J.W. Statistical Properties of Laser Speckle Patterns. In *Laser Speckle and Related Phenomena*; Dainty, J.C., Ed.; Springer: Berlin/Heidelberg, Germany, 1975; pp. 9–75\_2. [[CrossRef](#)]
35. Duncan, D.D.; Kirkpatrick, S.J. Algorithms for simulation of speckle (laser and otherwise). In *Complex Dynamics and Fluctuations in Biomedical Photonics V*; Tuchin, V.V., Wang, L.V., Eds.; International Society for Optics and Photonics, SPIE: Bellingham, WA, USA, 2008; Volume 6855, pp. 23–30. [[CrossRef](#)]
36. Khaksari, K.; Kirkpatrick, S. Laser Speckle Modeling and Simulation for Biophysical Dynamics: Influence of Sample Statistics. *J. Biomed. Photonics Eng.* **2017**, *3*, 040302. [[CrossRef](#)]
37. Cennini, C. *The Book of the Art of Cennino Cennini: A Contemporary Practical Treatise on Quattrocento Painting*; Taylor & Francis: London, UK, 2018.
38. Fotakis, C.; Anglos, D.; Zafiroopoulos, V.; Georgiou, S.; Tornari, V. *Lasers in the Preservation of Cultural Heritage: Principles and Applications*; CRC Press: Boca Raton, FL, USA, 2006.
39. International Digital Image Correlation Society. *A Good Practices Guide for Digital Image Correlation*; Jones, E., Iadicola, M., Eds.; 2018. Available online: <https://idics.org/guide/> (accessed on 4 November 2020).
40. Kirkpatrick, S.; Duncan, D.; Wells-Gray, E. Detrimental effects of speckle-pixel size matching in laser speckle contrast imaging. *Opt. Lett.* **2009**, *33*, 2886–2888. [[CrossRef](#)]

41. Reu, P.L.; Sweat, W.; Miller, T.; Fleming, D. Camera System Resolution and its Influence on Digital Image Correlation. *Exp. Mech.* **2015**, *55*, 9–25. [[CrossRef](#)]
42. Pan, B.; Xie, H.; Wang, Z.; Qian, K.; Wang, Z. Study on subset size selection in digital image correlation for speckle patterns. *Opt. Express* **2008**, *16*, 7037–7048. [[CrossRef](#)]

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).





Article

# Camera Color Correction for Cultural Heritage Preservation Based on Clustered Data

Marco Trombini <sup>1</sup>, Federica Ferraro <sup>1</sup>, Emanuela Manfredi <sup>2</sup>, Giovanni Petrillo <sup>2</sup> and Silvana Dellepiane <sup>1,\*</sup>

<sup>1</sup> Department of Electrical, Electronics and Telecommunication Engineering and Naval Architecture, Università degli Studi di Genova, Via All’Opera Pia 11A, 16145 Genoa, Italy; marco.trombini@edu.unige.it (M.T.); federica.ferraro@edu.unige.it (F.F.)

<sup>2</sup> Department of Chemistry and Industrial Chemistry, Università degli Studi di Genova, Via Dodecaneso 31, 16146 Genoa, Italy; giselle1861@yahoo.it (E.M.); giovanni.petrillo@unige.it (G.P.)

\* Correspondence: silvana.dellepiane@unige.it; Tel.: +39-348-7920633

**Abstract:** Cultural heritage preservation is a crucial topic for our society. When dealing with fine art, color is a primary feature that encompasses much information related to the artwork’s conservation status and to the pigments’ composition. As an alternative to more sophisticated devices, the analysis and identification of color pigments may be addressed via a digital camera, i.e., a non-invasive, inexpensive, and portable tool for studying large surfaces. In the present study, we propose a new supervised approach to camera characterization based on clustered data in order to address the homoscedasticity of the acquired data. The experimental phase is conducted on a real pictorial dataset, where pigments are grouped according to their chromatic or chemical properties. The results show that such a procedure leads to better characterization with respect to state-of-the-art methods. In addition, the present study introduces a method to deal with organic pigments in a quantitative visual approach.

**Keywords:** color correction; chemical composition; camera characterization



**Citation:** Trombini, M.; Ferraro, F.; Manfredi, E.; Petrillo, G.; Dellepiane, S. Camera Color Correction for Cultural Heritage Preservation Based on Clustered Data. *J. Imaging* **2021**, *7*, 115. <https://doi.org/10.3390/jimaging7070115>

Academic Editors:

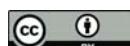
Giovanna Castellano, Gennaro Vessio and Fabio Bellavia

Received: 27 May 2021

Accepted: 10 July 2021

Published: 13 July 2021

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Cultural heritage bears witness to life and history, provides an identity to nations, and represents an irreplaceable source of inspiration. Its importance from cultural, historical, and economic points of view is invaluable; thus, its preservation and valorization are crucial topics for our society. Natural aging and deterioration due to external agents endanger artworks such as paintings, sculptures, and architecture, and therefore diagnostic tools are needed for monitoring and preservation.

Monitoring historical artistic heritage consists of the evaluation of possible modifications of some characteristics of the object under observation. When it comes to a artwork or, more generally, a mono- or polychromatic surface, color is one of those characteristics, as it is easily perceptible by the human eye, allows one to distinguish an artwork, and provides information on the nature and status of an artwork.

Color analysis on artworks is generally performed via specific instruments such as colorimeters and spectrophotometers, both of which use sophisticated technologies to accurately and precisely quantify and define color, working in a device-independent color space as Commission Internationale de l’Éclairage (CIE)  $L^*a^*b^*$  [1,2]. This allows for objective assessment of color changes in order to monitor the state of the painting over time and appropriately plan periodic protection or restoration actions. Color studies of artworks could also make use of Infrared (IR) and Ultraviolet (UV) data (by means of, e.g., infrared reflectography, UV–Visible spectrophotometry, UV reflectance, etc.) or X-ray fluorescence spectroscopy (XRF) [3,4].

However, several drawbacks may limit the efficacy of such devices/methodologies. First, colorimeters and spectrophotometers give, as with XRF, pointwise measurements;

thus, color studies on large areas require several time-consuming repetitions. Furthermore, even though spectrophotometers are defined as non-invasive devices [5], they must perfectly lean onto the artwork surface in order to exclude external light radiation, thus, risking ruining the painting. Finally, it is still rare nowadays that small laboratories are equipped with the abovementioned costly devices.

In order to address such issues, recent studies have proposed performing color monitoring through photographic documentation. Here, the necessary equipment is conceivably minimal and considerably cheaper, consisting of a professional digital camera and a photographic set with adequate lighting; then, the color data of each pixel of the selected area can be stored from a single photoshoot, limited only by illumination [6]. However, uncorrected digital data are not directly comparable, in terms of quantitative reliability, to the standard provided by the more specific spectrophotometric instrumentation. In addition, a well-defined procedure consisting of camera calibration, arrangement of lights, and positioning of the artwork, although necessary, is not sufficient per se for a correct comparison of digital data with colorimetric data. Finally, another major problem when using a digital camera for measuring color is that consumer-level sensors (either CCD or CMOS type) are typically uncalibrated.

Therefore, camera characterization is needed, i.e., some specific digital image processing to transform raw color digital values into objective  $L^*a^*b^*$  values equivalent to colorimetric measures.

A common approach to minimize the difference between digital and colorimetric determinations relies on the application of a correction based on a least-squares regression to the uncalibrated digital data. Linear [7,8], nonlinear, and mixed [9] approaches have all been described in the literature. Regarding nonlinear regressions, one can mention polynomial regressions [10,11], neural networks (NNs) [12,13], and look-up tables [14]. In addition, the problem of different color spaces based on the acquisition device must be addressed. Indeed, camera data usually refer to RGB or sRGB color spaces. Several approaches have been proposed [15], such as linear or quadratic models, neural networks for  $L^*a^*b^*$  regression starting from RGB values, and models requiring RGB data to be converted into XYZ values, which are then used to derive  $L^*a^*b^*$  values, with and without a linearization of sRGB data via a gamma model. Gamma correction is also involved in the method [16].

Further aiming at minimization of the correction error, other features to be preserved may be considered. For instance, characterization should be robust across different illuminants and reflectance types, and across noise [17–19].

To achieve better results, the use of digital image processing techniques for camera characterization can also be combined with different disciplines. Indeed, a multidisciplinary approach allows one to deal with specific features related to the heterogeneity of the data under analysis. Therefore, in order to overcome the lack of homoscedasticity required to apply a single-step procedure, an innovative approach combining pattern recognition and image processing techniques with chemistry information is proposed here.

In the present study, 117 tiles from the database of diagnostic analyses of The Foundation Centre for Conservation and Restoration of Cultural Heritage “La Venaria Reale” (in collaboration with the National Institute of Metrological Research and Laboratorio Analisi Scientifiche della Regione Autonoma Valle d’Aosta) represent the basic dataset [3,20].

As proposed by the state-of-the-art literature, the methods of linear regression, polynomial regression, and NN [8,9,19] were initially applied herein to the whole dataset, but the resulting performances unfortunately proved to be not satisfactory. It is worthwhile mentioning that a preliminary camera calibration using the X-Rite ColorChecker Passport Photo failed to provide satisfactory results [9], as expected, due to the limited color content of such a color chart.

To understand the reason for such poor results, the work was adapted by conducting a closer investigation of the pigments’ characteristics and their corresponding statistical analysis in photographic images in order to overcome the significant lack of the homoscedas-

ticity feature that is required for proper application of approaches in the literature, which work at a global level.

Consequently, from the perspective of optimizing the analysis, the original idea proposed herein is to apply state-of-the-art characterization methods to clusters of data rather than to the whole digital dataset, selected by means of two different criteria, i.e., the color and chemical properties of pigments. Regarding the latter, based on Kremer code, the main chemical element can be objectively defined for each pictorial layer analyzed.

To overcome the issue of a small amount of data and to find one-to-one correspondence between an image and colorimetric data, samples referring to the same tile are sorted by hue values, which provides coupled data and the use of supervised methods for precise and punctual color correction.

Thus, the application of several methods for camera characterization to numerous clusters of the base dataset is described hereinafter, in order to minimize the difference between digital and spectrophotometric quantitative color data, and therefore validate a handy diagnostic tool such as a digital camera for color determination. The best characterization approach results were achieved from a polynomial regression, while the predominant factor that affects the efficacy of the color correction could be found in the chemical composition, more precisely, in the nature of the central element. The best results were those splitting the data by chemical composition. In addition, the proposed method also proved to be effective with organic pigments, which could not be analyzed via standard approaches such as XRF; in fact, the latter has been employed to identify the presence of inorganic pigments, characterized by elements with an atomic number higher than 13. Instead, other non-invasive approaches for the study of organic pigments (usually referred to as “lakes”) include IR and Raman spectroscopy, but still require rather sophisticated instrumentation.

The considered approaches are briefly presented in Section 2, along with the dataset. Additionally, details on how data were collected and split into clusters and how camera images were used are provided.

Although a complete color analysis of artworks is also based on IR and UV data, the scope of the present study is to investigate how deep an analysis performed with traditional photographic data can be.

## 2. Materials and Methods

### 2.1. Background

Sensors' responses to light distribution are clearly defined in the literature [21,22]. For the sake of clarity, let  $I(\lambda)$  be the illuminant spectral power distribution falling on the surface patch ( $\lambda$  is the wavelength), and let  $\gamma(\lambda)$  be the reflectance function of the material the object is made from (or that its surface is painted with), so that the spectral power distribution  $P(\lambda)$  can be expressed as follows:

$$P(\lambda) = I(\lambda)\gamma(\lambda) \quad (1)$$

where  $P(\lambda)$  is the spectrum of the light that reaches the sensor and is associated with the corresponding pixels of the image.

Then, let  $\sigma(\lambda)$  be the spectral filter function of the sensor, and define the sensor's response to  $P(\lambda)$  as follows:

$$s = \int_{\lambda} P(\lambda)\sigma(\lambda)d\lambda \quad (2)$$

As mentioned in the Introduction, in the present study, camera and colorimeter sensors are involved. Hence, hereinafter, whenever  $s$  refers to the camera, it will be referred to as  $c_m$ , while the colorimeter,  $s$ , will be referred to as  $c_l$  (which will be the reference measurement). Specifically, based on the available data,  $c_m$  is a three-dimensional vector  $c_m = [R, G, B]^T$ , laying in the RGB color space. Similarly, the colorimeter response comes from the device-independent color space CIE L\*a\*b\*, namely  $c_l = [L, a, b]^T$ .

In order to perform an efficient correction on error-prone measurements of color changing, such as those deriving from commercial cameras, an optimal transformation  $f$  such that  $c_m \xrightarrow{f} c_l$  must be found. In fact, the final value of such a correction is only an approximation of the real corresponding  $c_m$  value, namely  $f(c_m) = \hat{c}_l$ , due to the different nature of the considered color spaces, to noise, estimation, and computation errors, etc. Some constraints can be added to improve the precision of the correction and are discussed later in the paper. The general requirement for the function  $f$  is to be error-minimizing, i.e.:

$$f = \operatorname{argmin}_g \sum_{i=1}^N \left\| u(c_l) - u(g(c_m)) \right\|, \quad (3)$$

where  $N$  is the number of color triplets in the dataset,  $u$  is a color space transformation to ensure that  $g(c_m)$  and  $c_l$  refer to the same color space, and  $\|\cdot\|$  is the norm. In the present manuscript, the considered norms will be the root mean squared error (Euclidean distance) and  $\Delta E_{00}$  [23]. In addition, a similarity measure will also be involved, i.e., Pearson's correlation coefficient.

In the following, since  $f$  is properly designed to correct  $c_m$  to be more similar to  $c_l$ , hence, both  $c_l$  and  $f(c_m)$  are in the CIE L\*a\*b\* color space, and  $u$  is assumed to be the identity function.

In general, methods in the literature are applied to the entire dataset. However, it appeared that no conditions for a single correction were present because of the non-homoscedasticity of the data. Hence, the methods were applied to clusters of tiles that could be determined according to some criterion. Here, this multi-cluster approach is based on either the chemical element or color, which is the major novelty of this study.

In this paper, the dependance of the color on the predominant chemical composition as well as on its chromaticity is investigated. More specifically, let  $C_i$  be the  $i$ -th cluster of color, based on either the chemical properties or the chromaticity. The purpose is to find many functions  $f_i$ , one for each cluster, which, of course, depends on the cluster that the input color belongs to:

$$c_m \in C_i \implies \hat{c}_l = f_i(c_m) = f(c_m | C_i) \quad (4)$$

## 2.2. Instrumentation

According to the CIE standard definition [24], reference measurements were made using a Konica Minolta CM2600d spectrophotometer (Konica Minolta, Ramsey, NJ, USA) [25] with the following setup: standard observer at  $10^\circ$ , illuminant D65, and acquisition SCI. Five measurements were acquired for each pictorial layer.

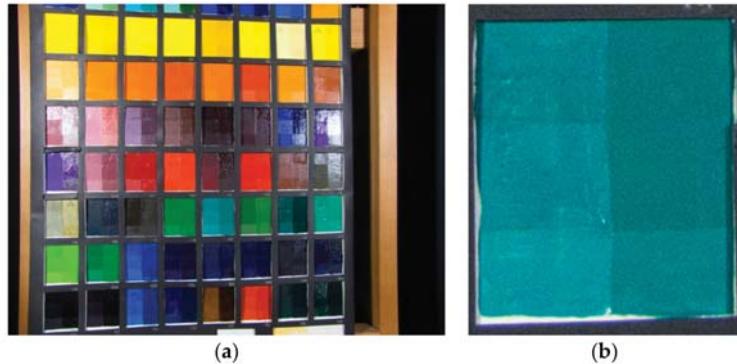
Photographic image data were acquired with a Lumix DMC-FZ200 camera (Panasonic, Osaka, Japan). The following image acquisition setup was used: The camera was placed vertically at 46.5 cm from the samples. The angle between the axis of the lens and the sources of illumination was approximately  $45^\circ$ . Illumination was achieved with two Natural Daylight 23 W fluorescent lights (OSRAM, Munich, Germany), color temperature 6500 K, reproducing the standard D65 illuminant. The photos were shot in a dark room. The settings of the camera are summarized in Table 1.

**Table 1.** Camera setup.

Variable	Value
Focal distance	4 mm
Flash	Off
ISO speed	400
Operation mode	Manual
Exposure time	1/60 s
Quality	Raw
f-Number	f/3.2

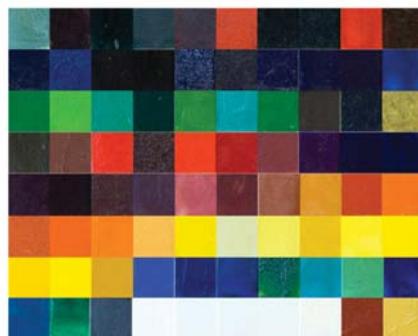
### 2.3. Dataset

As previously mentioned, the dataset of the present study consisted of 117 tiles from the database of diagnostic analyses of La Venaria Reale [20]. A picture for each tile was taken to enable analysis. Figure 1a shows an example of a photographic picture of the tables from Venaria.



**Figure 1.** (a) On the left, a picture of a table with a collection of colored tiles from The Foundation Centre for Conservation and Restoration of Cultural Heritage “La Venaria Reale”; (b) on the right, a single tile. The reader may notice the presence of two columns and three rows. The red box indicates the area considered for this study.

In the table, each pigment (Figure 1b) is presented in a mixture with two binders: polyvinyl acetate (PVAc) (column on the left) and linseed oil (column on the right). Then, the painted surface is divided into 3 rows. The first two present 2 different finishings: terpene resin (stripe on the top) and acrylic resin (middle stripe), while the third one is unprotected. For the present study, only the unprotected and the linseed oil sectors were taken into consideration (the red box in Figure 1b), because the linseed oil technique is the one most used by painters since the 15th century. The central portion of the camera acquisition was considered in order to avoid specularity and saturation problems. Figure 2 shows some of the selected parts of tiles involved in the study.



**Figure 2.** A subset of tile samples involved in the present study.

To address the local color inhomogeneity of tiles, the characterization was performed by taking into consideration five measurements via the colorimeter and five RGB triplets extracted from the pictures in order to create paired couples and to develop a robust supervised color correction.

Specifically, pixels from each tile were sorted by hue (in ascending order). Then, five triplets were extracted, namely the first one (i.e., the one with minimal hue), the last one (i.e., the one with maximal hue), and the ones corresponding to the 25th, 50th, and 75th percentiles. This was done to obtain as many samples as possible for the reference dataset.

#### 2.4. Linear Regression

The method consists of estimating the L\*, a\*, and b\* values separately via linear regression. In particular, let  $(\alpha_L, \beta_L)$ ,  $(\alpha_a, \beta_a)$ , and  $(\alpha_b, \beta_b)$  be the regression coefficients for L\*, a\*, and b\*, respectively, so that the estimated values are  $L_e = \alpha_L L_l + \beta_L$ ,  $a_e = \alpha_a a_l + \beta_a$ , and  $b_e = \alpha_b b_l + \beta_b$ , where  $L_l$ ,  $a_l$ , and  $b_l$  are the colorimeter values. To find the best characterization of the camera data with  $\widehat{L}_m$ ,  $\widehat{a}_m$ , and  $\widehat{b}_m$ , the constraints  $L_m \rightarrow \widehat{L}_m \cong L_l$ ,  $a_m \rightarrow \widehat{a}_m \cong a_l$ , and  $b_m \rightarrow \widehat{b}_m \cong b_l$  are added, yielding the following:

$$\widehat{L}_m = \frac{L_e - \beta_L}{\alpha_L}, \quad \widehat{a}_m = \frac{a_e - \beta_a}{\alpha_a}, \quad \widehat{b}_m = \frac{b_e - \beta_b}{\alpha_b} \quad (5)$$

#### 2.5. Polynomial Regression

The polynomial regression approach consists of mapping a polynomial expansion of the device RGB values to estimated L\*a\*b\*. In the following, the polynomial  $P_8$  was used:

$$P_8 = [R, G, B, RG, RB, GB, RGB, 1], \quad (6)$$

The corrected L\*a\*b\* triplet  $\hat{c}_l$  is obtained via the following equation:

$$\hat{c}_l = MP_8, \quad (7)$$

where  $M$  is the  $3 \times 8$  transformation matrix, which is derived via a pseudo-inversion procedure as in [11].

#### 2.6. Hue-Plane-Preserving Camera Characterization—Weighted Constrained Matrixing Method

The Hue-Plane-Preserving Camera Characterization—Weighted Constrained Matrixing (HPPCC-WCM) method [19] is aimed at ensuring that the characterization preserves the hue plane and minimizes error. Starting from the camera data, the transformation matrix is defined in function of the device hue angle  $\varphi_m$  and of the parameter  $p$  referring to the order of the transformation, as follows:

$$M(\varphi_m, p) = \frac{1}{\sigma} \sum_{i=1}^N (\pi - \Delta\varphi_i)^p M_i, \quad (8)$$

where  $N$  is the number of training coupled colorimeter–camera data  $(c_l, c_m)$ ,  $M_i$  is the transforming matrix  $c_{m,i} = M_i c_{l,i}$ ,  $\Delta\varphi_i = \min(|\varphi_m - \varphi_i|, 2\pi - |\varphi_m - \varphi_i|)$ , with  $\varphi_i$  being the  $i$ -th training color hue angle, and  $\sigma = \sum_{i=1}^N (\pi - \Delta\varphi_i)^p$ .

To sum up, the color correction here proposed is as follows:

$$\hat{c}_l = M(\varphi_m, p) c_m \quad (9)$$

#### 2.7. Data Grouping

To avoid the application of each method in a global way, the dataset under analysis was clustered according to two different criteria, i.e., according to chromatic appearance and the chemical composition, with reference to the central metal atom. Table A1 in Appendix A shows the available pigments and relevant features (pigment name and color, chemical composition, chemical cluster, and chromatic cluster). Regarding the chromatic appearance, five classes were subjectively identified. Conversely, regarding the chemical composition, Kremer code [26] was objectively considered. The clusters and relevant numbers of tiles are summarized in Table 2.

**Table 2.** Number of tiles for each selected cluster.

Red	33	Iron	23
Green	16	Lead	10
Blue	31	Copper	16
Yellow	23	Copper (organic)	9
Gray	14	Organic dyes and salts	34
		Iron, manganese, and cobalt	31
		Other	36

Some considerations regarding this clustering are made in the following.

In general, three phases drove the choice of the different chemical clusters. Firstly, three major classes were considered referring to the elements most spread in the dataset: iron, lead, and copper (Phase 1).

Then, by looking at the copper class, it was found that some tiles were organic lakes, generating the idea that this clustering method could also be effectively applied to organic dyestuff. Accordingly, the clusters “copper (organic)” and “organic” (collecting all the lakes in the dataset) were considered (Phase 2).

Finally, a mixed class was also considered, characterized by the presence of either iron, manganese, or cobalt, i.e., vicinal transition metals with very similar electronic properties (Phase 3).

Regarding the chromatic clusters, the gray cluster collects pigments with similar R, G, and B values (thus also including black and white pigments).

In Table A1, one can notice that the color grouping of some pigments differs from the chromatic class to which they belong, according to the closest color perception. For example, tile number 57, despite being visually brown/violet, also has shades of red given by its chemical description provided by Kremer, which identifies it as a red pigment.

### 2.8. Proposed Method

The proposed approach involves a combination of the aforementioned procedures. The data grouping procedure splits the dataset into clusters, which are homogeneous in terms of either color or chemical properties. The color correction methods are independently applied to each cluster. Recall that colorimetric and camera data are precisely coupled by hue, as specified in Section 2.3. Altogether, this leads to an adaptive color correction method.

## 3. Results

The color correction process was assessed via a five-fold cross-validation approach. The effectiveness of the procedure was evaluated on the grounds of statistical parameters such as Pearson’s correlation coefficient and the three measures of color distance. The root mean squared error in the L\*, a\*, and b\* parameters (RMS) and the related color distance measure expressed in color units, according to the formula  $\Delta = \sqrt{RMS(L)^2 + RMS(a)^2 + RMS(b)^2}$  [23], represent traditional metrics. The  $\Delta E_{00}$  distance, officially adopted in 2001 as the new CIE color difference equation, improves the performance on blue and gray colors thanks to an interactive term between chroma and hue differences and a scaling factor for the CIELAB a\* scale, respectively [27]. The latter is implemented, here, according to the CIEDE2000 formula [21] (MATLAB implementation [28]).

The relevant values in Tables 3–13 are the means of the five attempts performed during the cross-validation. In addition, the error associated with each value is specified in brackets. It was computed as the semi-difference between the maximum and minimum values in the five measurements, and it assesses the robustness of the k-fold procedure.

**Table 3.** Evaluation of the considered methods for the whole dataset. The bold font highlights the best values throughout.

	Pearson's Coefficient			RMS			
	L	a	b	L	a	b	
Uncalibrated	0.95 (0.03)	0.85 (0.02)	0.95 (0.02)	13.52 (3.94)	13.26 (4.28)	15.22 (5.12)	
Linear regression	0.80 (0.03)	−0.16 (0.01)	−0.34 (0.02)	106.70 (9.44)	126.18 (18.11)	98.22 (11.26)	
Polynomial regression	<b>0.95 (0.02)</b>	<b>0.91 (0.01)</b>	<b>0.96 (0.02)</b>	<b>8.17 (1.57)</b>	<b>9.89 (2.08)</b>	<b>10.53 (1.88)</b>	
HPPCC-WCM	0.94 (0.03)	0.87 (0.03)	0.44 (0.02)	38.30 (9.44)	46.83 (8.11)	90.74 (11.91)	
$\Delta E_{00}$			$\Delta$				
Uncalibrated	127.31 (10.5)			24.30 (4.16)			
Linear regression	140.13 (14.16)			192.23 (14.10)			
Polynomial regression	101.26 (8.93)			<b>16.60 (2.94)</b>			
HPPCC-WCM	<b>53.52 (7.18)</b>			109.06 (9.87)			

**Table 4.** Pearson's correlation coefficients of the considered methods for the major chemical clusters (Phase 1). The bold font highlights the best values throughout.

	Lead			Iron			
	L	a	b	L	a	b	
Uncalibrated	0.92 (0.03)	0.93 (0.03)	<b>0.97 (0.02)</b>	0.76 (0.02)	0.92 (0.02)	0.90 (0.02)	
Linear regression	0.59 (0.03)	−0.67 (0.03)	−0.87 (0.03)	0.87 (0.03)	0.16 (0.03)	−0.28 (0.02)	
Polynomial regression	<b>0.92 (0.02)</b>	<b>0.98 (0.02)</b>	<b>0.97 (0.02)</b>	<b>0.91 (0.02)</b>	<b>0.93 (0.02)</b>	<b>0.94 (0.01)</b>	
HPPCC-WCM	0.66 (0.02)	0.90 (0.03)	0.84 (0.01)	0.85 (0.03)	0.76 (0.03)	0.25 (0.02)	
Copper							
	L	a	b				
	0.89 (0.03)	0.51 (0.03)	0.78 (0.03)				
Uncalibrated	0.24 (0.01)	−0.16 (0.02)	0.02 (0.03)				
Linear regression	<b>0.91 (0.01)</b>	<b>0.87 (0.03)</b>	<b>0.92 (0.01)</b>				
Polynomial regression	0.66 (0.03)	0.84 (0.02)	0.81 (0.01)				

**Table 5.** Pearson's correlation coefficients of the considered methods for the other chemical clusters (Phase 2 and Phase 3). The bold font highlights the best values throughout.

	Copper (Organic)			Organic				
	L	a	b	L	a	b		
Uncalibrated	−0.31 (0.02)	0.75 (0.04)	0.66 (0.03)	0.91 (0.03)	0.88 (0.02)	0.90 (0.02)		
Linear regression	−0.52 (0.02)	−0.04 (0.01)	−0.52 (0.03)	0.89 (0.03)	−0.11 (0.04)	−0.32 (0.03)		
Polynomial regression	<b>0.77 (0.03)</b>	<b>0.90 (0.02)</b>	<b>0.92 (0.02)</b>	<b>0.97 (0.02)</b>	<b>0.91 (0.04)</b>	<b>0.92 (0.03)</b>		
HPPCC-WCM	0.23 (0.02)	−0.43 (0.04)	0.22 (0.03)	0.92 (0.01)	0.78 (0.03)	−0.53 (0.05)		
Iron + Mn + Co								
	L	a	b					
	0.77 (0.04)	0.93 (0.03)	0.90 (0.03)					
Uncalibrated	0.44 (0.03)	−0.56 (0.03)	−0.39 (0.02)					
Linear regression	<b>0.82 (0.02)</b>	<b>0.93 (0.02)</b>	<b>0.91 (0.02)</b>					
Polynomial regression	0.65 (0.03)	0.59 (0.03)	0.09 (0.03)					
HPPCC-WCM								

**Table 6.** RMS of the considered methods for the major chemical clusters (Phase 1). The bold font highlights the best values throughout.

	Lead			Iron		
	L	a	b	L	a	b
Uncalibrated	12.46 (2.49)	9.96 (2.33)	13.87 (2.92)	14.02 (1.91)	5.98 (1.4)	13.91 (1.74)
Linear regression	156.97 (9.88)	198.51 (8.91)	143.58 (8.56)	81.13 (6.5)	58.64 (5.19)	51.34 (7.92)
Polynomial regression	<b>5.29 (2.03)</b>	<b>4.25 (2.94)</b>	<b>6.89 (0.8)</b>	<b>5.22 (2.53)</b>	<b>4.04 (1.97)</b>	<b>4.89 (1.69)</b>
HPPCC-WCM	35.59 (7.37)	44.91 (6.39)	123.48 (5.89)	82.81 (6.95)	122.67 (5.48)	152.47 (5.76)
Copper						
	L	a	b			
Uncalibrated	12.42 (2.98)	19.45 (3.89)	14.97 (3.98)			
Linear regression	39.11 (7.13)	136.14 (6.89)	899.47 (44.13)			
Polynomial regression	<b>5.21 (1.42)</b>	<b>8.79 (3.03)</b>	<b>5.92 (1.78)</b>			
HPPCC-WCM	147.46 (14.71)	91.33 (6.86)	116.78 (12.03)			

**Table 7.** RMS of the considered methods for the other chemical clusters (Phase 2 and Phase 3). The bold font highlights the best values throughout.

	Copper (Organic)			Organic		
	L	a	b	L	a	b
Uncalibrated	13.07 (3.04)	10.44 (3.29)	12.02 (4.21)	17.77 (4.28)	23.12 (5.82)	21.38 (6.92)
Linear regression	31.48 (6.92)	166.89 (22.12)	66.34 (2.96)	164.24 (12.98)	231.84 (12.95)	89.29 (5.98)
Polynomial egression	<b>3.38 (0.24)</b>	<b>9.92 (2.31)</b>	<b>12.34 (4.32)</b>	<b>6.14 (3.07)</b>	<b>4.96 (0.45)</b>	<b>10.33 (1.89)</b>
HPPCC-WCM	142.29 (14.56)	113.95 (8.22)	125.13 (16.21)	39.34 (5.89)	87.43 (18.19)	194.90 (23.67)
Iron + Mn + Co						
	L	a	b			
Uncalibrated	15.72 (2.68)	7.44 (2.03)	14.49 (2.23)			
Linear regression	73.15 (9.86)	79.74 (9.65)	73.92 (6.73)			
Polynomial regression	<b>7.47 (2.33)</b>	<b>6.83 (1.12)</b>	<b>7.39 (2.11)</b>			
HPPCC-WCM	89.09 (7)	109.32 (13.13)	132.52 (16.18)			

**Table 8.**  $\Delta E_{00}$  of the considered methods for the chemical clusters. The bold font highlights the best values throughout.

	Lead	Iron	Copper	Copper (Organic)	Organic	Iron + Mn + Co
Uncalibrated	47.35 (9.68)	152.59 (13.72)	336.47 (24.68)	129.96 (7.34)	166.31 (29.76)	138.61 (12.91)
Linear regression	136.43 (13.8)	135.49 (14.25)	149.02 (18.32)	176.22 (16.43)	172.15 (32.94)	142.90 (19.94)
Polynomial regression	<b>9.78 (3.18)</b>	126.50 (10.18)	95.24 (8.58)	111.38 (13.23)	113.29 (22.63)	93.38 (8.22)
HPPCC-WCM	46.87 (5.97)	<b>79.42 (6.38)</b>	<b>68.48 (3.69)</b>	<b>82.11 (9.56)</b>	<b>78.62 (9.35)</b>	<b>83.09 (5.63)</b>

**Table 9.** Pearson's correlation coefficients of the considered methods for the chromatic clusters. The bold font highlights the best values throughout.

	Red			Green		
	L	a	b	L	a	b
Uncalibrated	0.86 (0.03)	0.90 (0.03)	0.92 (0.02)	0.89 (0.03)	0.70 (0.02)	0.84 (0.02)
Linear regression	0.86 (0.04)	0.13 (0.03)	-0.60 (0.03)	0.32 (0.03)	-0.35 (0.04)	-0.15 (0.02)
Polynomial regression	<b>0.92 (0.02)</b>	<b>0.91 (0.01)</b>	<b>0.94 (0.03)</b>	<b>0.91 (0.01)</b>	<b>0.84 (0.02)</b>	<b>0.86 (0.01)</b>
HPPCC-WCM	0.91 (0.02)	0.47 (0.02)	-0.30 (0.01)	0.54 (0.03)	0.83 (0.02)	0.74 (0.01)

Table 9. Cont.

	Blue			Yellow		
	L	a	b	L	a	b
Uncalibrated	0.89 (0.03)	0.38 (0.03)	0.84 (0.02)	0.87 (0.02)	0.35 (0.03)	0.81 (0.02)
Linear regression	0.74 (0.02)	−0.33 (0.02)	0.04 (0.02)	0.39 (0.02)	−0.43 (0.01)	−0.71 (0.03)
Polynomial regression	<b>0.92 (0.02)</b>	0.73 (0.02)	<b>0.89 (0.01)</b>	<b>0.90 (0.03)</b>	<b>0.87 (0.03)</b>	<b>0.92 (0.02)</b>
HPPCC-WCM	0.73 (0.02)	<b>0.87 (0.01)</b>	0.75 (0.02)	0.78 (0.01)	0.65 (0.01)	−0.13 (0.02)
	Gray					
	L	a	b			
Uncalibrated	0.99 (0.01)	0.74 (0.01)	0.76 (0.01)			
Linear regression	0.98 (0.02)	0.18 (0.03)	0.10 (0.02)			
Polynomial regression	<b>0.99 (0.01)</b>	0.88 (0.01)	0.75 (0.01)			
HPPCC-WCM	0.98 (0.02)	<b>0.99 (0.01)</b>	<b>0.95 (0.02)</b>			

Table 10. RMS of the considered methods for the chromatic clusters. The bold font highlights the best values throughout.

	Red			Green		
	L	a	b	L	a	b
Uncalibrated	14.72 (2.58)	9.70 (1.56)	14.68 (2.01)	13.58 (2.09)	14.32 (3.23)	12.77 (2.98)
Linear regression	138.58 (7.55)	62.36 (5.79)	53.43 (6.69)	47.86 (7.36)	142.85 (11.81)	80.63 (6.40)
Polynomial regression	<b>6.64 (1.73)</b>	<b>6.31 (0.82)</b>	<b>9.42 (1.72)</b>	<b>6.72 (1.23)</b>	<b>10.28 (2.34)</b>	<b>8.23 (0.82)</b>
HPPCC-WCM	36.15 (2.68)	129.90 (7.76)	176.77 (11.24)	118.15 (15.73)	69.96 (4.61)	104.65 (18.03)
	Blue			Yellow		
	L	a	b	L	a	b
Uncalibrated	13.61 (2.71)	14.22 (2.94)	14.65 (2.20)	12.90 (1.96)	17.31 (2.51)	19.75 (3.31)
Linear regression	49.17 (4.73)	78.28 (7.31)	310.77 (29.33)	163.72 (6.69)	204.02 (9.87)	83.26 (6.39)
Polynomial regression	<b>7.65 (0.74)</b>	<b>9.99 (2.06)</b>	<b>7.93 (1.05)</b>	<b>3.77 (1.48)</b>	<b>4.10 (1.24)</b>	<b>8.91 (1.61)</b>
HPPCC-WCM	60.56 (7.88)	49.43 (7.61)	51.42 (9.26)	14.48 (16.56)	65.89 (7.41)	252.45 (16.67)
	Gray					
	L	a	b			
Uncalibrated	8.98 (1.95)	9.48 (1.09)	15.71 (3.96)			
Linear regression	117.32 (9.34)	257.94 (8.92)	180.36 (8.27)			
Polynomial regression	<b>5.67 (1.42)</b>	<b>8.39 (2.04)</b>	<b>8.14 (1.89)</b>			
HPPCC-WCM	22.82 (4.22)	23.65 (3.92)	52.87 (4.54)			

Table 11.  $\Delta E_{00}$  of the considered methods for the chromatic clusters. The bold font highlights the best values throughout.

	Red	Green	Blue	Yellow	Gray
Uncalibrated	55.45 (6.21)	211.41 (22.39)	201.16 (19.35)	18.74 (4.56)	45.29 (6.86)
Linear regression	110.13 (7.47)	161.11 (5.69)	129.01 (8.07)	153.30 (8.51)	146.29 (6.94)
Polynomial regression	<b>35.12 (3.67)</b>	558.33 (44.10)	186.45 (13.96)	<b>5.83 (0.67)</b>	61.26 (5.90)
HPPCC-WCM	69.11 (7.23)	<b>68.03 (5.50)</b>	<b>79.12 (6.16)</b>	56.71 (5.15)	51.81 (5.28)

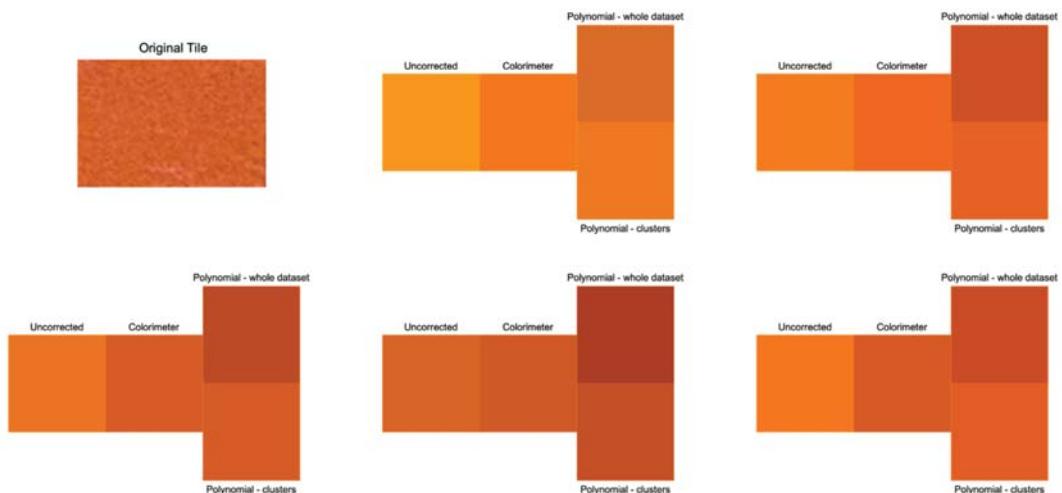
Table 12.  $\Delta$  of the considered methods for the chemical clusters. The bold font highlights the best values throughout.

	Lead	Iron	Copper	Copper (Organic)	Organic	Iron + Mn + Co
Uncalibrated	21.14 (4.82)	20.64 (5.96)	26.89 (5.17)	29.49 (5.25)	44.53 (11.72)	22.64 (5.75)
Linear regression	290.97 (13.67)	112.50 (8.02)	939.86 (22.37)	188.19 (9.89)	304.21 (35.28)	131.05 (10.07)
Polynomial regression	<b>9.67 (1.22)</b>	<b>8.21 (1.94)</b>	<b>11.32 (1.16)</b>	<b>22.77 (1.92)</b>	<b>19.18 (5.92)</b>	<b>12.53 (2.47)</b>
HPPCC-WCM	136.13 (13.64)	212.49 (12.88)	211.75 (16.36)	233.65 (9.37)	226.56 (18.37)	193.52 (9.53)

**Table 13.**  $\Delta$  of the considered methods for the chromatic clusters. The bold font highlights the best values throughout.

	Red	Green	Blue	Yellow	Gray
Uncalibrated	22.94 (4.35)	23.51 (5.01)	24.54 (4.02)	29.26 (5.19)	20.43 (5.25)
Linear regression	161.08 (13.49)	170.87 (9.13)	324.23 (18.08)	274.52 (13.35)	335.90 (16.27)
Polynomial regression	<b>13.14 (1.58)</b>	<b>14.78 (2.37)</b>	<b>14.87 (2.79)</b>	<b>10.51 (2.03)</b>	<b>12.99 (2.49)</b>
HPPCC-WCM	222.33 (13.35)	172.64 (8.35)	93.57 (8.26)	261.31 (11.31)	62.25 (5.78)

Some examples of color correction applied to the pigments are reported in Figure 3. All five measurements extracted from the considered tile are shown, coupled according to the described approach. Each visualization depicts the uncorrected color values, the colorimeter data, and the correction when the polynomial regression characterization was trained on the whole dataset and on the specific cluster. The reader may notice the improvement in the visual rendering when dealing with clustered data by chromatic properties.

**Figure 3.** Images depicting an example of color correction by means of the polynomial regression method.

#### 4. Discussion

First, the obtained results are discussed in terms of the values of the metric considered. Then, the importance of the preliminary cluster analysis is highlighted, with observations mainly relevant to the two clustering procedures. To conclude, possible applications of the proposed pipeline are disclosed, along with the limitations of the present research and foreseeable future developments.

##### 4.1. Discussing the Considered Indexes' Values

Table 3, referring to the application of the methods to the whole dataset, shows a strong agreement among the traditional metrics of correlation, the RMS on the  $L^*$ ,  $a^*$ , and  $b^*$  parameters, and the color distance  $\Delta$ .

In general, for both the whole dataset and the different selected clusters, the characterization method that produced the best color correction was polynomial regression, which was always able to improve similarity with colorimetric data as compared with uncalibrated data. Linear regression dramatically worsened the result as compared with the original data, as did the HPPCC-WCM method on most of the indexes. However, even though polynomial regression always showed improvements, according to the  $\Delta E_{00}$  distance, the

HPPCC-WCM method outperformed the others since both the method and the metric rely on the more recent CIE standards, facing some drawbacks of the traditional standards.

Tables 4–7 confirm the best performances of polynomial regression, which improved uncalibrated data on all clusters, even in the challenging case of copper (organic), where the acquired colorimetric and photographic L parameters showed a strong misalignment.

Table 8 gives further evidence that the  $\Delta E_{00}$  metric can solve some problems of traditional colorimetry as, except for the lead cluster, it gives better improvements. Additional results reported in Table 11 show that the blue, green, and, to a lesser extent, the gray clusters might benefit from hue preservation and the new metric, as declared in the new standard scope.

The Pearson coefficients were already high for the whole dataset; hence, the improvement obtained by clustering was less relevant for this index. Conversely, by taking into consideration RMS,  $\Delta E_{00}$ , and  $\Delta$ , the improvement when passing from the global correction to the cluster-based correction was significant, as they both decreased when focusing on chemical and chromatic clusters. In fact, the expression of the prediction error in terms of color units is only intended to evaluate the human perception of the correction; indeed, recall that if the error is approximately less than 2.2 color units, then, the difference is considered to be imperceptible to the human eye. It is worthwhile noting the improvement in this index, which decreased from a mean value across classes of 27.56 (Table 12, “uncalibrated”) to 13.95 with respect to the chemical clusters (Table 12, “polynomial regression”), and from a mean value across classes of 24.14 (Table 13, “uncalibrated”) to 13.26 with respect to the chromatic clusters (Table 13, “polynomial regression”). In such a case, clustering based on the chemical components is the most effective procedure, i.e., the one producing the lowest error. It is expected that more sophisticated algorithms, which could be investigated in future developments of the present study, would lead to an even lower color unit error.

#### 4.2. The Significance of the Clustering Procedure

Splitting the dataset into clusters led to a better color correction for both splitting criteria (chromatism or chemical composition). The efficacy of clustering can be appreciated by comparing the value of, for example, the  $\Delta$  index for the whole dataset (Table 3, 16.60 after polynomial regression characterization, with a 32% decrease with respect to the value for the uncalibrated data) with the values for the single clusters in Tables 12 and 13, for example, for the “lead” cluster, the value is 9.67, with a 54% decrease after characterization. Therefore, one can infer that the clustering procedure effectively addresses the homoscedasticity of the data. Indeed, the major contribution of the present study is the efficiency of the coupling between clustering and application of some state-of-the-art color correction methods. In addition, it is worth stressing that, although one might expect better results and a more effective color correction from chromatic clusters, the best correction was provided by chemical clusters. This is likely due to the objectivity of the chemical component criterion for defining clusters, while chromatic properties are more dependent on human perception, thus, leading to less homogeneous classes.

To stress once more the novelty of the present study, to the best of our knowledge, such an approach as well as the results relevant to the different clustering criteria are unprecedented.

#### 4.3. Chromatic Clusters

As outlined above, clustering by chromatism seems less effective for color correction purposes. The perceived colors driving the selection of the chromatic clustering depend, to some extent, on the observer, and therefore are subjective. In addition, several color shades are present, which may lead to heterogeneous classes. Of course, having more samples for each tile would allow one to split the data into more classes, each being characterized by a closer chromatic similarity; as a result, the training phase would benefit, thus, conceivably leading to better color correction.

The correction provided by the polynomial regression method on the coordinates of the Lab color space suggests some additional considerations, recalling that “L” represents the perceptual lightness, while “a” and “b” refer to the four colors in the opposite component model of human vision, i.e., red, green, blue, and yellow.

The most improved coordinate was L, meaning that this procedure addresses the problems in terms of the lightness sensitivity of photographic data. Without correction, the error is so high that the observer perceives a consistently different color with respect to the colorimetric data (see Figure 3).

Regarding the coordinates “a” and “b”, it is interesting to consider the chromatic class of gray. The values in this class are supposed to be similar, and so the difference between colorimeter and photographic data should also be similar. However, by looking at the RMS index (Table 10), we found that the difference between colorimeter and photographic data was much higher for “b” than for “a”. Conversely, once the values were corrected with polynomial regression, the errors were similar, thus, suggesting that the procedure is useful to address some imbalance for the gray class.

#### 4.4. Chemical Clusters

The criterion based on chemical composition is more univocal, an aspect that surely contributes, in general, to the results being more similar as well as rewarding across the considered classes.

In particular, the RMS index mirrors a gratifying, effective color correction for the main elemental clusters (lead, iron, and copper, see Table 6) after polynomial regression characterization.

Regarding the additional elemental classes reported in Table 2, particular attention must be paid to the copper-based samples. In fact, the “copper” cluster of Table 2 (16 samples) also included nine organic samples, where copper was the metal cation of an organic salt, which made up the selected subcluster defined as “copper (organic)”. In terms of the RMS index, both the “copper” cluster and the subcluster performed extremely well (Tables 6 and 7, respectively) as far as the “L” component was concerned, while the components “a” and “b” did not seem to be significantly corrected for the subcluster by the characterization method of choice. Nonetheless, we paid attention to the consistent number of tiles containing organic pictorial matter (either organic dyes or metal salts of organic acids); satisfactorily enough, the rather crowded (34 samples of lakes) “organic dyes and salts” cluster responded positively to the polynomial regression characterization (as compared with the values of the RMS index in Table 7 or of the  $\Delta$  index in Table 12) or to the HPPCC-WCM treatment (as compared with the value of the  $\Delta E_{00}$  index in Table 8).

The performance provided by the cluster of lakes represents, in our opinion, a further original and very interesting aspect of the camera characterization procedure herein. This is because the identification and study of organic matter on pictorial artworks cannot be achieved by means of XRF, a non-invasive technique that is widely applied in the presence of pigments containing heavy metals, but which fails to detect organic dyestuff because C, N, and O atoms are too light. Instead, the current approach based on the correction of digital data grouped in elemental clusters does not depend on the atomic weight, and thus opens very appealing perspectives as to the analysis of lakes. Developments and applications to study cases are necessary to sustain this hypothesis.

A first hypothesis about the reason why elemental clustering is a good approach is suggested by the results of the analysis on the mixed “iron, manganese, and cobalt” cluster. In fact, these three elements are transition metals adjacent in the periodic table, whose electronic configuration differs only for the number of electrons at the internal level, with the external one being identical for all three. The good results obtained for such a mixed class may mean that the proposed approach is sensitive to the outermost electronic level. Of course, more experimental trials are needed to validate the hypothesis, particularly by selecting other mixed clusters responding to the same characteristics.

#### 4.5. A Possible Usage of the System for the Programming of Restoration Actions

A main concern about cultural heritage is the preservation of artwork for future generations. Of course, artworks, whatever the typology, inevitably tend to change or degrade with time due to several different causes, and restoration campaigns must be conducted whenever necessary. As far as pictorial artworks are concerned, color is surely the main sentinel to be observed in order to decide what actions to take. A handy and low-cost tool such as a digital camera would be optimal for frequent periodic control on artworks, as well as on large surfaces. In this way, time-dependent data describing the state of the paintings could be easily collected and analyzed preliminarily to further deepen more sophisticated analyses, if necessary, prior to a restoration action.

To this end, repeated periodical collections of data are necessary to verify the feasibility of selecting a parameter as a valid index of color deterioration. While elemental clustering has proven optimal for the identification of color, it could be foreseen that chromatic clustering would be best to handle the fading/deterioration of color with time. Of course, at the present time, this is only a conjecture to be verified in the future as a compulsory development of the present study.

#### 4.6. Limitations and Future Developments

First, the amount of available data needs to be increased, as it is supposed that it would lead to better correction, at least on statistical grounds.

Of course, the present study cannot be limited to “theory”; in addition to the desirable significance of the method outlined in the previous paragraph, a main interest would be the application of the training to real cases in order to perform identification and study of the pictorial layers of an unknown composition. Thus, once the chemical clusters have been characterized, one can consider an “unknown” painting and focus on a particular area. If such an area fits a particular cluster, i.e., proper color correction is obtained by considering the parameters for that class, then it would mean that the relevant chemical elements are present in the considered area.

A continuing collaboration with the laboratories of “La Venaria Reale” and contacts with museums or galleries would surely satisfy both the outlined forms of progress and enable the development of a novel machine-learning-based approach, which is presently hampered by the limited size of the available dataset.

### 5. Conclusions

A dataset of digital camera photographs and of colorimetric measurements on 117 tiles from the database of diagnostic analyses of The Foundation Centre for Conservation and Restoration of Cultural Heritage “La Venaria Reale” was collected and analyzed with the aim of minimizing the difference between digital and spectrophotometric quantitative color data, from the perspective of validating a handy diagnostic tool such as a digital camera for quantitative color determination.

To address the homoscedasticity of the data acquired, the current study proposed a supervised approach to camera characterization and color correction based on clustered data. To this end, within the dataset, samples were grouped into clusters based on either the chromatic or the chemical properties of the pigments.

Among the different approaches studied in the present study, a polynomial regression obtained the best results with both of the proposed clustering criteria. Thus, while the correlation between characterized photographic data and colorimetric data remains high when considering both the entire dataset and the single clusters, in the latter case, notable improvements can be seen in the three parameters considered to test the efficacy of the characterization (i.e., RMS,  $\Delta E_{00}$ , and  $\Delta$ ). The central thesis that the piecewise method improves prediction accuracy was supported by numerical evaluations, even though, in absolute terms, the results were short of an error low enough to be imperceptible to a human expert.

In future studies, the aim could be to extend the dataset, for example, by developing the collaboration with La Venaria Reale. Of course, increasing the dataset would allow one to define new or more densely populated clusters, and therefore study the chemical and chromatic properties of the pigments in more detail, hopefully confirming the hypotheses above. A larger dataset may substantially improve the error, and therefore achieve imperceptible differences between the acquired data and the corrected data.

Furthermore, different approaches could be investigated, no longer based on the mean value of the colorimetric data, but rather looking for other significant parameters to perform the analysis. Additionally, further applications of the proposed approach are being investigated, such as applying it for characterizing the chemical composition of unknown artworks by leveraging the photographic data.

**Author Contributions:** Conceptualization, G.P. and S.D.; methodology, M.T. and F.F.; software, M.T. and F.F.; validation, M.T., F.F., G.P. and S.D.; formal analysis, M.T. and F.F.; investigation, M.T., F.F., G.P. and S.D.; resources, E.M.; data curation, M.T., F.F., E.M., G.P. and S.D.; writing—original draft preparation, M.T. and F.F.; writing—review and editing, G.P. and S.D.; visualization, M.T. and F.F.; supervision, G.P. and S.D.; project administration, G.P. and S.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors would like to thank Marco Nervo and Tiziana Cavaleri (Foundation Centre for Conservation and Restoration of Cultural Heritage “La Venaria Reale”), who made the present study possible by providing access to databases of colorimetric analyses.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

In this appendix, a table presenting the pigments' descriptions is provided.

**Table A1.** List of the available pigments with their name, chemical composition, and chemical and chromatic classes.

Tile	Pigment Name	Pigment Color	Chemical Composition	Chemical Class	Chromatic Class
Tile 1	Lead White	Lead white	2PbCO <sub>3</sub> .Pb(OH) <sub>2</sub>	Lead	Gray
Tile 4	Calcium Carbonate	White	CaCO <sub>3</sub>	Other	Gray
Tile 5	Vine Black German	Black	Retouching color in aldehyde resin 81. Carbon with impurities of potassium and sodium ions	Organic	Gray
Tile 9	Azurite MP, sky-blue light	Azure	Cu <sub>3</sub> (CO <sub>3</sub> ) <sub>2</sub> .(OH) <sub>2</sub>	Copper	Blue
Tile 13	Smalt	Blue	K <sub>2</sub> O.nSiO <sub>2</sub> with the presence of cobalt	Cobalt	Blue
Tile 15	Orpiment	Yellow	As <sub>2</sub> S <sub>6</sub>	Other	Yellow
Tile 16	Realgar	Orange, yellow	As <sub>4</sub> S <sub>4</sub>	Other	Yellow
Tile 19	Yellow Ochre Iron Oxide	Yellow	α-FeO(OH) or K <sub>2</sub> Fe(SO <sub>4</sub> ) <sub>2</sub> (OH) <sub>6</sub> or γ-FeO(OH))	Iron	Yellow
Tile 20	Raw Sienna Italian	Raw sienna, yellow, brown	Fe <sub>2</sub> O <sub>3</sub> .nH <sub>2</sub> O + MnO <sub>2</sub> + Al <sub>2</sub> O <sub>3</sub> + SiO <sub>2</sub>	Iron	Yellow
Tile 21	Massicot Litharge	Yellow	PbO	Lead	Yellow

Table A1. Cont.

Tile	Pigment Name	Pigment Color	Chemical Composition	Chemical Class	Chromatic Class
Tile 22	Burnt Sienna Italian	Burnt sienna, red, brown	$\text{Fe}_2\text{O}_3 \cdot \text{nH}_2\text{O} + \text{Al}_2\text{O}_3$ (60%) + $\text{MnO}_2$ (1%)	Iron	Red
Tile 23	Burnt Umber Reddish	Burnt umber, brown	$\text{Fe}_2\text{O}_3 + \text{MnO}_2 + \text{Si} + \text{Al}_2\text{O}_3$	Iron	Red
Tile 25	French Ochre SOFOROUGE	Red	$\text{SiO}_2 + \text{Al}_2\text{O}_3 + \text{Fe}_2\text{O}_3$	Iron	Red
Tile 28	Pozzuolana Red Earth	Purple, red	Mix of lands	Iron	Red
Tile 30	Madder Lake Genuine	Pink, red	Organic nature	Organic	Red
Tile 31	Red Ochre English	Red	$\text{Fe}_2\text{O}_3 \cdot \text{nH}_2\text{O}$	Iron	Red
Tile 33	Red Bole	Brown, red	$\text{Al}_2\text{Si}_2\text{O}_5(\text{OH})_4$	Other	Red
Tile 34	Red Lead, minimum	Orange, red	$\text{PbO}_4$	Lead	Red
Tile 36	Malachite Natural Standard	Green	$\text{Cu}_3(\text{CO}_3)_2 \cdot (\text{OH})_2$	Copper	Green
Tile 41	Verdigris, synthetic	Blue, turquoise, green	$\text{Cu}(\text{CH}_3\text{COO})_2$	Copper (organic)	Blue
Tile 42	Barium Sulfate	White	$\text{BaSO}_4$	Other	Gray
Tile 43	Sepia Fine	Black-brown	Sepia, fine (colorant of cuttlefish)	Organic	Gray
Tile 44	Bone Black	Black	15–20% of carbon, 60–70% of $\text{Ca}_3(\text{PO}_4)_2$	Other	Gray
Tile 45	Asphaltum Black	Black	High molecular weight hydrocarbons	Organic	Gray
Tile 46	Blue Bice	Blue, turquoise	$\text{Cu}_2(\text{CO}_3)_2 \cdot \text{Cu}(\text{OH})_2$	Copper	Blue
Tile 47	Lapis Lazuli	Blue	$(\text{Na},\text{Ca})_8(\text{AlSiO}_4)_6 + \%$ of iron	Iron	Blue
Tile 48	Ultramarine Ash	Blue ultramarine	$\text{Na}_2\text{O}_3\text{Al}_6\text{SiO}_2 \cdot 2\text{Na}_2\text{S}$	Other	Blue
Tile 49	Lead Tin Yellow Light	Lemon yellow	Lead stannate, type I ( $\text{Pb}_2\text{SnO}_4$ )	Lead	Yellow
Tile 50	Indian Yellow Imitation	Indian yellow	Consisting primarily of euxanthic acid salts	Organic	Yellow
Tile 51	Naples Yellow, dark	Naples yellow	$\text{Pb}_2\text{Sb}_2\text{O}_7$	Lead	Yellow
Tile 52	Van Dyck Brown	Van Dyke brown	Consists mainly of humic acids and iron oxide	Iron	Red
Tile 53	Natural Cinnabar	Orange, red	Mineral cinnabar, $\text{HgS}$	Other	Red
Tile 54	Lac Dye	Pink, red	Lac dye (from <i>coccus lacta</i> secretion, Natural Red 25; gum lac, Indian lake)	Organic	Red
Tile 55	Vermilion	Vermilion red	Mine cinnabar, $\text{HgS}$	Other	Red
Tile 56	Caput Mortuum Reddish	Red, violet	$\text{Fe}_2\text{O}_3$	Iron	Red
Tile 57	Caput Mortuum Violet	Brown, violet	$\text{Fe}_2\text{O}_3$	Iron	Red
Tile 58	Green Earth Light	Green	Iron-based silicate	Iron	Green
Tile 59	Prussian Blue	Prussian blue	$\text{Fe}_4[\text{Fe}(\text{CN})_6]_3 \cdot 6\text{H}_2\text{O}$ or $\text{KFe}[\text{Fe}(\text{CN})_6] \cdot 6\text{H}_2\text{O}$	Iron	Blue
Tile 60	Lead Tin Yellow II	Lemon yellow	Type II, $\text{Pb}(\text{Sn},\text{Si})\text{O}_3$	Lead	Yellow

Table A1. Cont.

Tile	Pigment Name	Pigment Color	Chemical Composition	Chemical Class	Chromatic Class
Tile 61	Naples Yellow from Paris	Yellow	Pb(Sb,Sn)O <sub>3</sub>	Lead	Yellow
Tile 62	Venetian Red	Venetian red	Fe <sub>2</sub> O <sub>3</sub>	Iron	Red
Tile 67	Lead Sulfate	White	PbSO <sub>4</sub>	Lead	Gray
Tile 68	Lithopone	White	BaSO <sub>4</sub> + ZnS	Other	Gray
Tile 69	Titanium White Rutile	Titanium white	TiO <sub>2</sub>	Other	Gray
Tile 70	Zinc White	Zinc white	ZnO + % of iron	Iron	Gray
Tile 71	Zinc Sulfide	White	ZnS	Other	Gray
Tile 72	Manganese Black	Black	(Fe,Mn) <sub>3</sub> O <sub>4</sub>	Iron	Gray
Tile 73	Ploss Blue	Blue, turquoise,	(CuCa)CO <sub>3</sub> (CH <sub>3</sub> COO) <sub>2</sub> .2H <sub>2</sub> O	Copper (organic)	Blue
Tile 74	Blue Verditer	Blue	CuCO <sub>3</sub> Cu(OH) <sub>2</sub>	Copper	Blue
Tile 75	Ultramarine Blue very dark	Ultramarine blue	Al <sub>6</sub> Na <sub>8</sub> O <sub>24</sub> S <sub>3</sub> Si <sub>6</sub>	Other	Blue
Tile 76	Copper Blue	Blue, turquoise	Copper based	Copper	Blue
Tile 77	Zirconium cerulean blue	Cerulean blue	Derived from zircon	Other	Blue
Tile 78	Cavansite	Blue, turquoise	Ca(VO)Si <sub>4</sub> O <sub>10.4</sub> (H <sub>2</sub> O)	Other	Blue
Tile 79	Ultramarine Blue Dark	Ultramarine blue	Na <sub>2</sub> O <sub>3</sub> Al <sub>6</sub> SiO <sub>2</sub> .2Na <sub>2</sub> S	Other	Blue
Tile 80	Cobalt Blue Dark	Blue	(Co,Zn) <sub>2</sub> SiO <sub>4</sub>	Cobalt	Blue
Tile 81	Cobalt Blue Pale	Blue	CoAl <sub>2</sub> O <sub>4</sub>	Other	Blue
Tile 82	Natural Chromium Yellow or crocoite	Yellow	PbCrO <sub>4</sub>	Lead	Yellow
Tile 83	Cadmium Yellow n°6 medium	Cadmium yellow	CdS + ZnO	Other	Yellow
Tile 84	Permanent Yellow medium	Yellow	Organic nature	Organic	Yellow
Tile 85	Brilliant Yellow	Yellow	C <sub>18</sub> H <sub>18</sub> N <sub>4</sub> O <sub>6</sub>	Organic	Yellow
Tile 86	Studio Yellow	Yellow	C <sub>16</sub> H <sub>12</sub> Cl <sub>2</sub> N <sub>4</sub> O <sub>4</sub>	Organic	Yellow
Tile 87	Cobalt Yellow	Yellow	[Co(NO <sub>2</sub> ) <sub>6</sub> ]K <sub>3</sub> + 3H <sub>2</sub> O	Cobalt	Yellow
Tile 88	Bismuth-Vanadate YelLow Lemon	Yellow	(Bi,V)O <sub>4</sub>	Other	Yellow
Tile 89	Baryte Yellow	Yellow	BaCrO <sub>4</sub>	Other	Yellow
Tile 90	Studio Pigment Yellow	Yellow	C <sub>18</sub> H <sub>18</sub> N <sub>4</sub> O <sub>6</sub>	Organic	Yellow
Tile 91	Studio Pigment Yellow Sun Gold	Yellow	Organic nature	Organic	Yellow
Tile 92	Cadmium Orange n°0 very light	Orange	Cd <sub>2</sub> SSe	Other	Red
Tile 93	Paliotol® Orange	Orange	C <sub>8</sub> H <sub>9</sub> N	Organic	Red
Tile 94	Paliogen® Orange	Orange	C <sub>23</sub> H <sub>8</sub> C <sub>18</sub> N <sub>4</sub> O <sub>2</sub>	Organic	Red

Table A1. Cont.

Tile	Pigment Name	Pigment Color	Chemical Composition	Chemical Class	Chromatic Class
Tile 95	Irgazin® Yellow, light orange	Yellow	C <sub>8</sub> H <sub>7</sub> NO	Organic	Yellow
Tile 96	Isoindolol Orange	Orange	C <sub>8</sub> H <sub>7</sub> N	Organic	Red
Tile 97	Titanium Orange	Orange, yellow	Ti-Sb-Cr-O Rutile	Other	Yellow
Tile 98	Iron Oxide Orange 960	Orange	Fe(O)OH + Fe <sub>2</sub> O <sub>3</sub>	Iron	Red
Tile 99	IWA-Enogu® Shinsia	Pink	Sodium aluminosilicate with oxides of metals other than iron	Other	Red
Tile 100	IWA-Enogu® Iwamomo	Pink	Sodium aluminosilicate with oxides of metals other than iron	Other	Red
Tile 101	IWA-Enogu® Usukuchi-Murasaki	Violet	Sodium aluminosilicate with oxides of metals other than iron	Other	Red
Tile 102	Côte d'Azur Violet	Gray, violet	Fe <sub>2</sub> O <sub>3</sub>	Iron	Gray
Tile 103	Thioindigo Red Lightfast	Red	Organic nature	Organic	Red
Tile 104	Cinquasia® Violet RT 201 D	Reddish violet	Organic nature	Organic	Red
Tile 105	Ultramarine Violet medium	Violet, bluish	Sodium, alumino, sulfo, silicate	Other	Blue
Tile 106	Manganese Violet	Manganese violet	(NH <sub>4</sub> ) <sub>2</sub> Mn <sub>2</sub> (P <sub>2</sub> O <sub>7</sub> )	Manganese	Blue
Tile 107	Cobalt Violet Dark	Cobalt violet	Co <sub>3</sub> (PO <sub>4</sub> ) <sub>2</sub>	Cobalt	Blue
Tile 108	Pink color	Pink, red	Ca(Sn,Cr)SiO <sub>5</sub>	Other	Red
Tile 109	Cadmium Red n°2 medium	Cadmium red	CdS	Other	Red
Tile 110	Irgazine® Scarlet DPP EK	Scarlet red	C <sub>6</sub> H <sub>2</sub> N <sub>2</sub> O <sub>2</sub>	Organic	Red
Tile 111	Alizarine Crimson dark	Pink, red	Colorant/organic pigment for coatings	Organic	Red
Tile 112	XSL Irgazine® Red DPP	Red	Organic nature	Organic	Red
Tile 113	Rosso Sartorius	Brown, red	Fe <sub>2</sub> O <sub>3</sub> .nH <sub>2</sub> O	Iron	Red
Tile 114	Aegirine Fine	Green	NaFeSi <sub>2</sub> O <sub>6</sub>	Iron	Green
Tile 115	Andeer Green Fine	Green	Granite	Other	Green
Tile 116	Phthalo Green dark	Green, bluish	Cu(C <sub>32</sub> N <sub>8</sub> Cl <sub>14</sub> ).16HCl	Copper (organic)	Green
Tile 117	Chromite	Green	FeCr <sub>2</sub> O <sub>4</sub>	Iron	Green
Tile 118	Cobalt Green	Green	Co <sub>2</sub> SnO <sub>4</sub>	Cobalt	Green
Tile 119	Cobalt Green Bluish A	Green, turquoise	Cobalt-based	Cobalt	Green
Tile 120	Chrome Oxide Green	Green	Cr <sub>2</sub> O <sub>3</sub>	Other	Green
Tile 122	Permanent Green	Green, turquoise	CoAl <sub>2</sub> O <sub>4</sub>	Cobalt	Green
Tile 123	Cadmium Green Light	Green	Cadmium-based	Other	Green

Table A1. Cont.

Tile	Pigment Name	Pigment Color	Chemical Composition	Chemical Class	Chromatic Class
Tile 124	Cadmium Green Dark	Green	Cadmium-based	Other	Green
Tile 125	Fluorescent Pigment Blue	Blue	Unspecified	Other	Blue
Tile 126	Phthalo Blue	Blue	$C_{32}H_{16}CuN_8$	Copper (organic)	Blue
Tile 127	Phthalo Blue Royal Blue	Blue	$C_{32}H_{16}CuN_8$	Copper (organic)	Blue
Tile 129	Indanthren® Blue	Blue	Unspecified	Organic	Blue
Tile 130	XSL Phthalo Blue Royal Blue Very Lightfast	Blue	$C_{32}H_{16}CuN_8$	Copper (organic)	Blue
Tile 131	Indigo Blue Lake	Blue	Organic nature	Organic	Blue
Tile 132	Indigo Red-Violet	Blue, violet	Organic nature	Organic	Blue
Tile 133	Studio Pigment Sky Blue	Blue	Unspecified	Other	Blue
Tile 134	Studio Pigment Dark Blue	Blue	Unspecified	Other	Blue
Tile 135	XSL Translucent Yellow	Yellow	Unspecified	Iron	Yellow
Tile 136	IWA-Enogu® Iwabeni	Red	Unspecified	Other	Red
Tile 137	Phthalo Green, yellowish	green, yellowish	Copper-based	Copper (organic)	Green
Tile 138	Heliogen® Green	Green, bluish	Copper-based	Copper (organic)	Green
Tile A	Madder Lake glazing over natural cinnabar	Red	Organic nature	Organic	Red
Tile B	Azurite glazing over Madder Lake	Blue	$Cu_3(CO_3)_2 \cdot (OH)_2$	Copper	Blue
Tile C	Lapis Lazuli glazing over Azurite	Blue	$(Na,Ca)_8(AlSiO_4)_6 + \text{ % of iron}$	Iron	Blue
Tile D	Copper resinate glazing over Verdigris	Green	$Cu(CH_3CO)_2 \cdot 2Cu(OH)_2 \cdot nH_2O$	Copper (organic)	Green
Tile E	Bisso (mixture of Madder Lake and Lapis Lazuli)	Blue	Organic nature	Organic	Blue
Tile F	Mixture of Azurite and Lead Tin Yellow Light	Green	$Cu_3(CO_3)_2 \cdot (OH)_2$ and $Pb_2SnO_4$	Copper lead (*)	Green

(\*): Tile F presents both copper and lead in its chemical composition, so it belongs to both copper and lead clusters.

## References

1. Ibraheem, N.A.; Hasan, M.M.; Khan, R.Z.; Mishra, P.K. Understanding color models: A review. *ARPN J. Sci. Technol.* **2012**, *2*, 265–275.
2. Sanmartín, P.; Chorro, E.; Vázquez-Nion, D.; Martínez-Verdú, F.M.; Prieto, B. Conversion of a digital camera into a non-contact colorimeter for use in stone cultural heritage: The application case to Spanish granites. *Measurement* **2014**, *56*, 194–202. [[CrossRef](#)]
3. Cavalieri, T.; Buscaglia, P.; Migliorini, S.; Nervo, M.; Piccablotto, G.; Piccirillo, A.; Zucco, M. Pictorial materials database: 1200 combinations of pigments, dyes, binders and varnishes designed as a tool for heritage science and conservation. *Appl. Phys. A* **2017**, *123*, 419. [[CrossRef](#)]
4. Pelagotti, A.; Mastio, A.D.; Rosa, A.D.; Piva, A. Multispectral imaging of paintings. *IEEE Signal Process. Mag.* **2008**, *25*, 27–36. [[CrossRef](#)]
5. Brunetti, B.; Miliani, C.; Rosi, F.; Doherty, B.; Monico, L.; Romani, A.; Sgamellotti, A. Non-invasive investigations of paintings by portable instrumentation: The MOLAB experience. *Top. Curr. Chem.* **2016**, *374*, 10. [[CrossRef](#)] [[PubMed](#)]

6. Jackman, P.; Sun, D.W.; ElMasry, G. Robust colour calibration of an imaging system using a colour space transform and advanced regression modelling. *Meat Sci.* **2012**, *91*, 402–407. [[CrossRef](#)] [[PubMed](#)]
7. Johnson, T. Methods for characterizing colour scanners and digital cameras. *Displays* **1996**, *16*, 183–191. [[CrossRef](#)]
8. Vrhel, M.J. Mathematical Methods of Color Correction. Ph.D. Thesis, North Carolina State University, Raleigh, NC, USA, 1993.
9. Manfredi, E.; Petrillo, G.; Dellepiane, S. A Novel Digital-Camera Characterization Method for Pigment Identification in Cultural Heritage. In *International Workshop on Computational Color Imaging*; Springer: Cham, Switzerland, 2019; pp. 195–206.
10. Finlayson, G.D.; Drew, M.S. Constrained least-squares regression in color spaces. *J. Electron. Imaging* **1997**, *6*, 484–493. [[CrossRef](#)]
11. Hong, G.; Luo, M.R.; Rhodes, P.A. A Study of Digital Camera Colorimetric Characterization Based on Polynomial Modeling. *Color Res. Appl.* **2001**, *26*, 76–84. [[CrossRef](#)]
12. Cheung, T.L.V.; Westland, S. Color camera characterisation using artificial neural networks. In Color and Imaging Conference. *Soc. Imaging Sci. Technol.* **2002**, *2002*, 117–120.
13. Xinwu, L. A new color correction model for based on BP neural network. *Adv. Inf. Sci. Serv. Sci.* **2011**, *3*, 72–78.
14. Hung, P.C. Colorimetric calibration in electronic imaging devices using a look-up-table model and interpolations. *J. Electron. Imaging* **1993**, *2*, 53–61. [[CrossRef](#)]
15. Leon, K.; Mery, D.; Pedreschi, F.; Leon, J. Color measurement in L\* a\* b\* units from RGB digital images. *Food Res. Int.* **2006**, *39*, 1084–1091. [[CrossRef](#)]
16. Cheung, V.; Westland, S.; Thomson, M. Accurate estimation of the nonlinearity of input/output response for color cameras. *Color Res. Appl.* **2004**, *29*, 406–412. [[CrossRef](#)]
17. Vora, P.; Herley, C. Trade-Offs Between Color Saturation and Noise Sensitivity in Image Sensors. In Proceedings of the 1998 International Conference on Image Processing, Chicago, IL, USA, 7 October 1998; Cat. No. 98CB36269. Volume 1, pp. 196–200.
18. Finlayson, G.D.; Mackiewicz, M.; Hurlbert, A. Color correction using root-polynomial regression. *IEEE Trans. Image Process.* **2015**, *24*, 1460–1470. [[CrossRef](#)] [[PubMed](#)]
19. Andersen, C.F.; Connah, D. Weighted constrained hue-plane preserving camera characterization. *IEEE Trans. Image Process.* **2016**, *25*, 4329–4339. [[CrossRef](#)] [[PubMed](#)]
20. Pictorial Material Database. Available online: [https://webimgc.inrim.it/Hyperspectral\\_imaging/Database.aspx](https://webimgc.inrim.it/Hyperspectral_imaging/Database.aspx) (accessed on 16 February 2021).
21. Sharma, G. Color Fundamentals for Digital Imaging. In *Digital Color Imaging Handbook*; CRC Press: Boca Raton, FL, USA, 2003; Volume 20, pp. 1–114.
22. Petrou, M.M.; Petrou, C. *Image Processing: The Fundamentals*; John Wiley & Sons: Hoboken, NJ, USA, 2010.
23. Sharma, G.; Wu, W.; Dalal, E.N. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Res. Appl.* **2005**, *30*, 21–30. [[CrossRef](#)]
24. International Commission of Illumination. Available online: <http://cie.co.at> (accessed on 24 February 2021).
25. Konica Minolta CM2600d Documentation. Available online: [https://www.konicaminolta.com/instruments/download/catalog/color/pdf/cm2600d\\_catalog\\_eng.pdf](https://www.konicaminolta.com/instruments/download/catalog/color/pdf/cm2600d_catalog_eng.pdf) (accessed on 24 February 2021).
26. Kremer Pigmente. Available online: <https://shop.kremerpigments.com> (accessed on 24 February 2021).
27. Luo, M.R.; Cui, G.; Rigg, B. The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Res. Appl.* **2001**, *26*, 340–350. [[CrossRef](#)]
28. Qiu, J. Calculate Color Difference Delta E of a Set of Srgb Data Pairs. 2021. Available online: <https://www.mathworks.com/matlabcentral/fileexchange/53156-calculates-color-difference-delta-e-of-a-set-of-srgb-data-pairs> (accessed on 24 February 2021).



Article

# Restoration and Enhancement of Historical Stereo Photos <sup>†</sup>

Marco Fanfani <sup>1</sup>, Carlo Colombo <sup>1</sup>, and Fabio Bellavia <sup>2,\*</sup>

<sup>1</sup> Department of Information Engineering, Università degli Studi di Firenze, 50139 Firenze, Italy; marco.fanfani@unifi.it (M.F.); carlo.colombo@unifi.it (C.C.)

<sup>2</sup> Department of Math and Computer Science, Università degli Studi di Palermo, 90123 Palermo, Italy

\* Correspondence: fabio.bellavia@unipa.it; Tel.: +39-0912-389-1124

† This paper is an extended version of our paper published in The 25th International Conference on Pattern Recognition Workshop (ICPRW 2020) on Fine Art Pattern Extraction and Recognition (FAPER 2020), Online, 10–15 January 2021.

**Abstract:** Restoration of digital visual media acquired from repositories of historical photographic and cinematographic material is of key importance for the preservation, study and transmission of the legacy of past cultures to the coming generations. In this paper, a fully automatic approach to the digital restoration of historical stereo photographs is proposed, referred to as Stacked Median Restoration plus (SMR+). The approach exploits the content redundancy in stereo pairs for detecting and fixing scratches, dust, dirt spots and many other defects in the original images, as well as improving contrast and illumination. This is done by estimating the optical flow between the images, and using it to register one view onto the other both geometrically and photometrically. Restoration is then accomplished in three steps: (1) image fusion according to the stacked median operator, (2) low-resolution detail enhancement by guided supersampling, and (3) iterative visual consistency checking and refinement. Each step implements an original algorithm specifically designed for this work. The restored image is fully consistent with the original content, thus improving over the methods based on image hallucination. Comparative results on three different datasets of historical stereograms show the effectiveness of the proposed approach, and its superiority over single-image denoising and super-resolution methods. Results also show that the performance of the state-of-the-art single-image deep restoration network Bringing Old Photo Back to Life (BOPBtL) can be strongly improved when the input image is pre-processed by SMR+.



**Citation:** Fanfani, M.; Colombo, C.; Bellavia, F. Restoration and Enhancement of Historical Stereo Photos. *J. Imaging* **2021**, *7*, 103. <https://doi.org/10.3390/jimaging7070103>

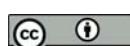
Academic Editor: Anna Tonazzini

Received: 20 April 2021

Accepted: 21 June 2021

Published: 24 June 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



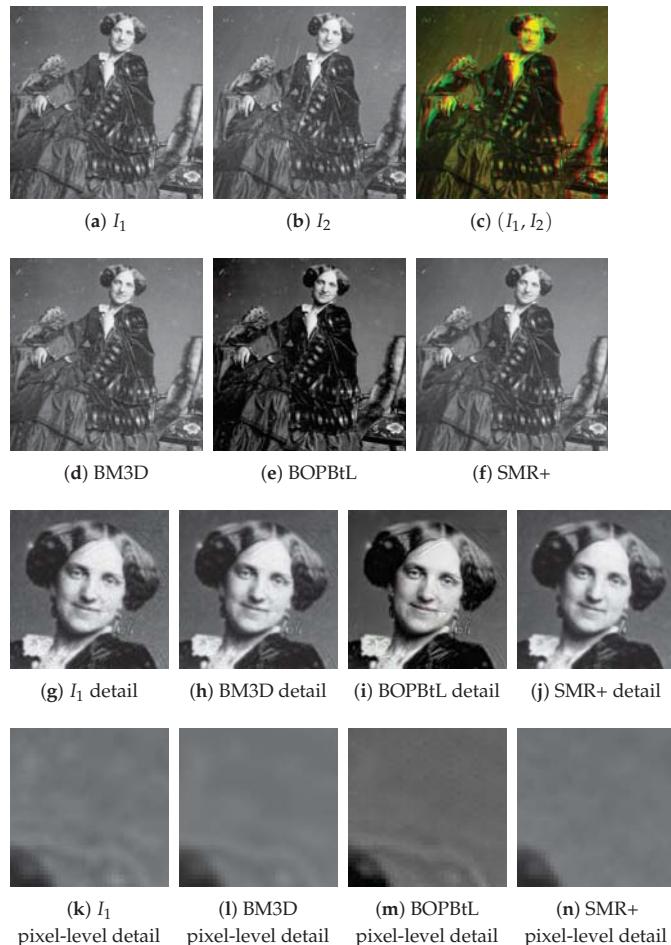
**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Photographic material of the XIX and XX centuries is an invaluable source of information for historians of art, architecture and sociology, as it allows them to track the changes occurred over the decades to a community and its living environment. Unfortunately, due to the effect of time and bad preservation conditions, most of the survived photographic heritage is partially damaged, and needs restoration, both at the physical (cardboard support, glass negatives, films, etc.) and digital (the image content acquired through scanners) levels. Dirt, scratches, discoloration and other signs of aging strongly reduce the visual quality of photos [1]. A similar situation also holds for the cinematographic material [2].

Digital restoration of both still images and videos has attracted considerable interest from the research community in the early 2000s. This has led to the development of several tools that improve the visual quality. Some approaches rely on the instantiation of noise models, which can either be fixed a priori or derived from the input images [3–5]. Other approaches detect damaged areas of the image and correct them according to inpainting techniques [6]. Self-correlation inside the image, or across different frames in videos, is often exploited in this context, under the assumption that zero-mean additive noise cancels out as the available number of image data samples increases [7–9]. A similar idea is exploited

by super-resolution techniques that enhance image quality by pixel interpolation [10,11]. In recent years, the algorithmic methods above have been sided by methods based on deep learning that can infer the image formation model or the scene content [12] from a training set in order to inject this information into the final output, a process called image hallucination [13–15]. Although the final image may often alter the original image data content, and hence cannot be fully trusted (e.g., in the medical diagnosis domain), the hallucination methods can give visually pleasing results (see Figure 1).



**Figure 1.** First row: An example of historical stereo pair of images,  $I_1$  and  $I_2$ , also superimposed as anaglyph. Second row: Enhancement of  $I_1$  according to different methods, including the proposed SMR+ method. Although visually impressive, the deep super-resolution result of BOPBtL does not preserve the true input image. Third row: A detail of  $I_1$  and the restored images according to the different methods. A closer look at BOPBtL reveals alterations with respect to the original face expression, accentuating the smile and introducing bush-like textures on the hair. Fourth row: pixel level detail of  $I_1$  and of the restored images according to the different methods. The specific image region considered is the background around the right shoulder. Notice the chessboard-like texture pattern typical of the deep network approaches, not visible at coarser scales. Best viewed in color: the reader is invited to zoom in on the electronic version of the manuscript in order to better appreciate the visual differences.

Stereoscopy has accompanied photography since its very birth in the nineteenth century, with ups and downs in popularity through time. Notwithstanding the lesser spread of stereo photography with respect to standard (monocular) photography, many digital archives with thousands of stereo images exist, some of which are freely available on the web. Stereo photos have a richer content than standard ones, as they present two different views of the same scene, thus explicitly introducing content redundancy and implicitly embedding information about scene depth. This characteristic can be exploited also in digital noise removal, enhancement and restoration, since a damaged area in one image can be reconstructed from the other image, provided that the correspondences between the two images are known. At a first glance, the above-mentioned approach looks similar to that of video restoration from multiple video frames, in which the scene is acquired in subsequent time instants from slightly changed viewpoints. However, stereo images have their own peculiarities, and actually introduce in the restoration process more complications than video frames, which in movies typically exhibit an almost static and undeformed background, differently from stereo pairs. As a matter of fact, although several advances have been recently made in stereo matching and dense optical flow estimation [16], the problem is hard and far from being fully solved, especially in the case of very noisy and altered images such as those generated by early photographic stereo material. To the best of the authors' knowledge, stereo photo characteristics have been employed only for the super-resolution enhancement or deblurring of modern, clean photos [17–19]. On the other hand, the image analysis and computer vision approaches developed so far for historical stereo photos mainly aimed at achieving (usually in a manual way) better visualizations or 3D scene reconstructions [20–22], with no attempt at restoring the quality of the raw stereo pairs.

This paper proposes a new approach to clean up and restore the true scene content in degraded historical stereo photographs, named Stacked Median Restoration Plus (SMR+), extending our previous work [23], and working in a fully automatic way. With respect to existing single image methods, damaged image areas with scratches or dust can be better detected and fixed, thanks to the availability of more sampled data points for denoising. In addition, the correct illumination can be restored or enhanced in a way akin to that of High Dynamic Range Imaging, where the images of the same scene taken at different exposure levels are used in order to enhance details and colors [24]. For this scope, the optical flow, estimated with the recent state-of-the-art Recurrent All-Pairs Field Transforms (RAFT) deep network [16], is used to synthesize corresponding scene viewpoints in the stereo pair, while denoising and restoration are carried out using novel yet non-deep image processing approaches. The entire process is superseded by scene content consistency validation, used to check critical stereo matching mispredictions that were left unresolved by the network. Our approach aims to obtain an output which is fully consistent with the original scenario captured by the stereo pair, in contrast with the recent super-resolution and denoising approaches based on image hallucination.

This paper extends our previous work [23], hereafter reported as Stacked Median Restoration (SMR) under several aspects:

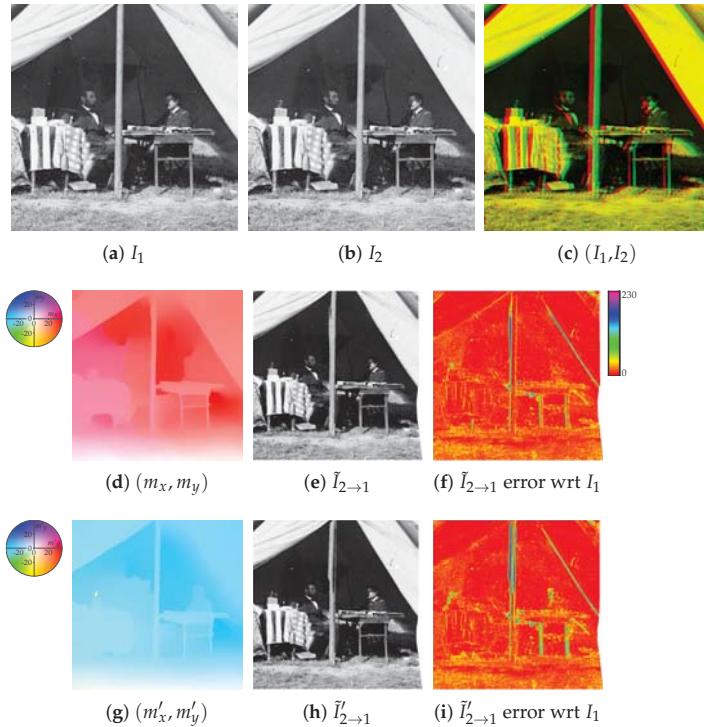
- With respect to SMR, the novel SMR+ is redesigned so as to better preserve finer details while at the same time improving further the restoration quality. This is accomplished by employing supersampling [25] at the image fusion step in conjunction with a weighting scheme guided by the original restoration approach.
- The recent state-of-the-art deep network BOPBL [26], specifically designed for old photo restoration, is now included in the comparison, both as standalone and to serve as post-processing of SMR+.
- The collection of historical stereo photos employed as a dataset is roughly doubled to provide a more comprehensive evaluation.
- The use of renowned image quality assessment metrics is investigated and discussed for these kinds of applications.

The rest of the paper is organized as follows: Section 2 introduces the proposed approach. An experimental evaluation and comparison with similar approaches is reported in Section 3. Finally, conclusions and future work are discussed in Section 4.

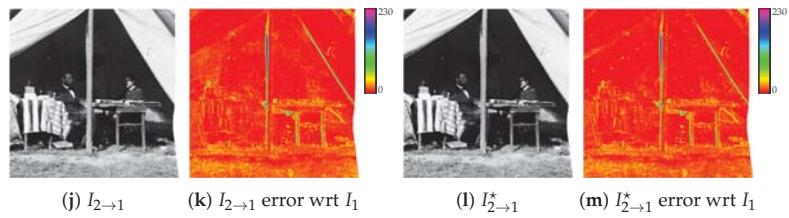
*Note: To ease the inspection and the comparison of the different images presented, an interactive PDF document is provided in the additional material (<https://drive.google.com/drive/folders/1Fmsm50bMMDSd0z4jXOhCZ3hPDIXdwMUL>) to allow readers to view each image at its full dimensions and quickly switch to the other images to be compared.*

## 2. Proposed Method

Given a pair of stereo images  $I_1$  and  $I_2$ , the aim of the process is to output a defect-free version of one image of the pair (referred to as the reference) by exploiting the additional information coming from the other image (denoted as auxiliary). For convenience, the reference is denoted as  $I_1$  (see Figure 2a) and the auxiliary image as  $I_2$  (see Figure 2b), but their roles can be interchanged. Images are assumed to be single channel graylevel, i.e.,  $I_1, I_2 : \mathbb{R}^2 \rightarrow [0, 255]$ .



**Figure 2.** *Cont.*



**Figure 2.** Auxiliary image pointwise transfer and color correction steps (see Sections 2.1 and 2.2): (a) Reference image  $I_1$ , (b) auxiliary image  $I_2$ , (c) superimposition of  $I_1$  and  $I_2$  as anaglyph, (d) visual representation of the optical flow map  $(m_x, m_y)$  extracted by RAFT, (e) image  $\tilde{I}_{2 \rightarrow 1}$  as resynthesis of  $I_2$  through  $(m_x, m_y)$  and (f) its error with respect to  $I_1$ , (g) visual representation of the optical flow map  $(m'_x, m'_y)$  extracted by RAFT after switching the input images, (h) image  $\tilde{I}'_{2 \rightarrow 1}$  as resynthesis from  $I_2$  through  $-(m'_x, m'_y)$  and (i) its error with respect to  $I_1$ , (j) final resynthesized image  $I_{2 \rightarrow 1}$  considering the locally best optical flow estimation between  $\tilde{I}_{2 \rightarrow 1}(x, y)$  and  $\tilde{I}'_{2 \rightarrow 1}(x, y)$  and (k) its error with respect to  $I_1$ , (l) image  $I_{2 \rightarrow 1}^*$  obtained after applying GPS/LCP color correction to  $I_{2 \rightarrow 1}$  using  $I_1$  as reference, and (m) the corresponding error map with respect to  $I_1$ . Best viewed in color. The reader is invited to zoom into the electronic version of the manuscript in order to better appreciate the visual differences.

### 2.1. Auxiliary Image Pointwise Transfer

As a first step, the recent state-of-the-art RAFT deep network [16] is used to compute the optical flow map pair  $f_{\text{RAFT}}(I_1, I_2) = (m_x, m_y)$ , where  $m_x, m_y : \mathbb{R}^2 \rightarrow \mathbb{R}$  (see Figure 2d), so that a synthesized image based on the content of  $I_2$  and registered onto  $I_1$  can be obtained as

$$\tilde{I}_{2 \rightarrow 1}(x, y) = I_2(x + m_x(x, y), y + m_y(x, y)) \quad (1)$$

by transferring pixel intensity values from  $I_2$  into the view given by  $I_1$  (see Figure 2e). Note that spots of missing data can be present on  $\tilde{I}_{2 \rightarrow 1}$  when no pixel in  $I_2$  maps onto the specific image area, due, for instance, to image occlusions. In the error free ideal case, it must hold that  $I_1 = \tilde{I}_{2 \rightarrow 1}$  for every correspondence between  $I_1$  and  $I_2$ . However, in real situations, this may not happen, as shown in Figure 2f reporting the average absolute error between  $I_1$  and  $\tilde{I}_{2 \rightarrow 1}$  on  $5 \times 5$  local window patches.

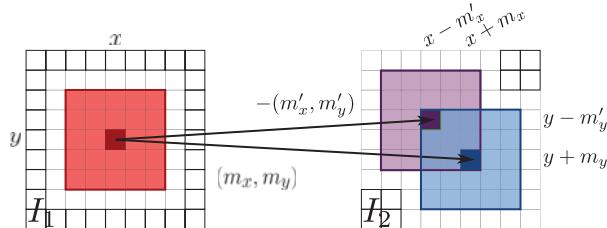
Notice also that, in the case of perfectly rectified stereo images, it holds everywhere that  $m_y(x, y) = 0$ . Under this particular setup, in which  $m_x$  is denoted as *disparity map* and is the only map that needs to be estimated, several classical methods have proven to provide good results while being computationally efficient [27]. However, according to our experience [21], these methods are not feasible in the case of degraded historical stereo photos. First, image degradation due to aging and the intrinsic image noise due to the technological limitations of the period decrease the ability of these methods to find the right correspondences. Second, the output of these methods is quite sensitive to the initial configuration of the parameters and, by considering the variability of the historical acquisition setups, each stereo pair would require the human supervision to get even a sub-optimal result. Third, the stereo alignment for the photos under consideration is far from perfect due to the technological limitations of the period, hence both the maps  $m_x$  and  $m_y$  are to be considered. Hence, our choice fell under the state-of-the-art RAFT that provides a sufficiently good initial estimation of the optical flow maps in most cases.

A further flow mapping pair  $f_{\text{RAFT}}(I_2, I_1) = (m'_x, m'_y)$  (see Figure 2g) can be obtained by switching the two input images, which can be employed to synthesize a second image according to

$$\tilde{I}'_{2 \rightarrow 1}(x, y) = I_2(x - m'_x(x, y), y - m'_y(x, y)) \quad (2)$$

(see Figure 2h) so that, in the error free ideal case for every correspondence between  $I_1$  and  $I_2$ , it holds that  $(m_x, m_y) = -(m'_x, m'_y)$ , which implies that  $I_1 = \tilde{I}_{2 \rightarrow 1} = \tilde{I}'_{2 \rightarrow 1}$ . This usually

does not happen, as shown by the relative error image of Figure 2i. Indeed, comparing the first and second rows of Figure 2, RAFT optical flow estimation is not completely accurate and does not preserve map inversion when exchanging the input image order. The final synthesized image  $I_{2 \rightarrow 1}$  (see Figure 2j) is then obtained by choosing the intensity value at each pixel  $(x, y)$  as the one from  $\tilde{I}_{2 \rightarrow 1}(x, y)$  and  $\tilde{I}'_{2 \rightarrow 1}(x, y)$  that minimizes the sum of absolute errors with respect to  $I_1$  on a small  $5 \times 5$  local window centered on the pixel (see Figure 3). A smaller error between the final resynthesized image  $I_{2 \rightarrow 1}$  and the reference image  $I_1$  is obtained (see Figure 2k) with respect to the errors given by  $\tilde{I}_{2 \rightarrow 1}(x, y)$  and  $\tilde{I}'_{2 \rightarrow 1}(x, y)$ .



**Figure 3.** Illustration of the  $I_{2 \rightarrow 1}$  image formation process from the two resynthesized images  $\tilde{I}_{2 \rightarrow 1}(x, y)$  and  $\tilde{I}'_{2 \rightarrow 1}(x, y)$ , respectively driven by the optical flow estimation maps  $(m_x, m_y)$  and  $-(m'_x, m'_y)$ . A point  $(x, y)$  in  $I_1$  can be mapped back to  $I_2$  according to either Equation (1) or Equation (2). The best back-mapping minimizing locally the error among the two possible optical flow estimates is then chosen to form  $I_{2 \rightarrow 1}$ . Best viewed in color.

## 2.2. Color Correction

Due to the technical limitations of the old photographic instrumentation, illumination conditions between the two stereo images can differ noticeably. For instance, flash lamp and, even more, flash powder did not provide each time uniform and identical illumination conditions, and it was not infrequent that a single camera was moved in two different positions in order to simulate a stereo setup instead of having two synchronized cameras [21]. Moreover, discoloration of the support due to aging can be present. In order to improve the final result, the state-of-the-art color correction method named Gradient Preserving Spline with Linear Color Propagation (GPS/LCP) presented in [28] is employed to correct the illumination of  $I_{2 \rightarrow 1}$  according to  $I_1$ . Specifically, the color map  $g_{\text{GPS/LCP}}(I_1, I_{2 \rightarrow 1}) = C$ , with  $C : \mathbb{R} \rightarrow \mathbb{R}$  is used to obtain the color corrected image  $I_{2 \rightarrow 1}^*$  according to

$$I_{2 \rightarrow 1}^*(x, y) = C(I_{2 \rightarrow 1}(x, y)) \quad (3)$$

where, in the error free ideal case, it must hold that  $I_1 = I_{2 \rightarrow 1}^*$  (see Figure 2l). The GPS/LCP color correction method is able to preserve the image content and works also in the case of not perfectly aligned images. Color correction decreases the resynthesis error. This can be noted by comparing the error map of  $I_{2 \rightarrow 1}^*$  (Figure 2m) with the error map of  $I_{2 \rightarrow 1}$  (Figure 2k), see for instance the error corresponding to the dark background above the left table. Clearly, if  $I_{2 \rightarrow 1}$  presents better illumination conditions than  $I_1$ , it is also possible to correct  $I_1$  according to  $I_{2 \rightarrow 1}$ .

## 2.3. Data Fusion

Given the reference image  $I_1$  and the synthesized one obtained from the auxiliary view  $I_{2 \rightarrow 1}^*$  after the illumination post-processing, the two images are blended into a new image  $I_{12}$  according to the *stacked median* operator (see Figure 4a)

$$I_{12} = \Theta(I_1 \cup I_{2 \rightarrow 1}^*) \quad (4)$$

The stacked median  $\Xi(\{I\})$  for a set of images  $\{I\}$  outputs a new image defined so that image intensity at pixel  $(x, y)$  is the median intensity value computed on the union of the pixels in the  $3 \times 3$  local neighbourhood windows centered at  $(x, y)$  on each image of the set (see Figure 5). Notice that the *median stacking* operation typically found as a blending tool in image manipulation software corresponds to the proposed stacked median operator with degenerate  $1 \times 1$  local windows. Unlike median stacking, the proposed definition does not require more than two input images and considers pixel neighborhoods, i.e., it works locally and not pointwise. Additionally, in case of missing data in  $I_{2 \rightarrow 1}^*$ , the stacked median acts as a standard  $3 \times 3$  median filter. With this operator, dirt, scratches and other signs of photographic age or damages are effectively removed from  $I_{12}$ , but high frequency details can be lost in the process, due to the  $3 \times 3$  filtering (see Figure 4b). These are partially re-introduced by considering a blended version of the gradient magnitude

$$d_{m_{12}} = \Xi(M(I_1) \cup M(I_{2 \rightarrow 1}^*)) \quad (5)$$

(see Figure 4c) obtained as the stacked median of eight possible gradient magnitudes, four for each of the  $I_1$  and  $I_{2 \rightarrow 1}^*$  images, to further enhance finer details. Each gradient magnitude image in the set  $M(I)$  for a generic image  $I$  is computed as

$$d_m = (d_x^2 + d_y^2)^{\frac{1}{2}} \quad (6)$$

pixelwise, where the image gradient direction pairs  $(d_x, d_y)$  are computed by the convolution of  $I$  with the following four pairs of kernel filters:

$$\left\{ \left( \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \right), \left( \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \right), \left( \begin{bmatrix} 0 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \right), \left( \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \right) \right\} \quad (7)$$

Notice that  $d_{m_{12}} \neq \Xi(M(I_{12}))$  in the general case (compare Figure 4c with Figure 4f). Consider for now only a single derivative pair  $(d_x, d_y)$  of  $I_{12}$ : Each pixel intensity  $I_{12}(x, y)$  is incremented by a value  $v(x, y)$  satisfying

$$\left( d_x + \frac{v}{2} \right)^2 + \left( d_y + \frac{v}{2} \right)^2 = \frac{d_m^2 - d_{m_{12}}^2}{2} \quad (8)$$

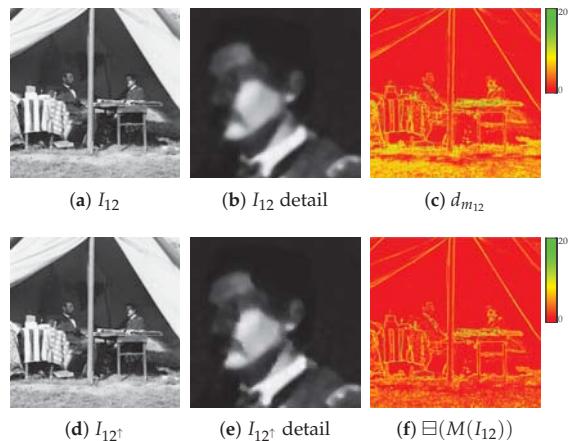
This equation has a twofold solution

$$v^* = \pm (2d_x d_y - d_{m_{12}}^2)^{\frac{1}{2}} - (d_x + d_y) \quad (9)$$

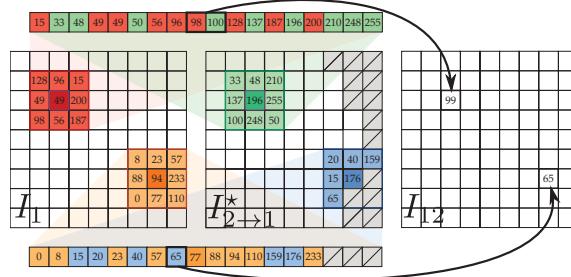
In the case of two real  $v^*$  solutions,  $v$  is chosen as  $v(x, y) = \arg \min_{\bar{v} \in v^*} |\bar{v}|$  in order to alter  $I_{12}$  as little as possible. In the case of complex solutions,  $v(x, y)$  is set to 0. The final gradient-enhanced image is then obtained as

$$I_{12}^\uparrow = I_{12} + v \quad (10)$$

(see Figure 4d,e for details). Since four different  $v$  values are obtained for each of the four derivative pairs of Equation (7), their average value is actually employed.



**Figure 4.** Data fusion step (see Section 2.3): (a) stacked median  $I_{12}$  obtained from  $I_1 \cup I_{2 \rightarrow 1}^*$ , (b) details of  $I_{12}$ , (c) the stacked median  $d_{m_{12}}$  of the gradient magnitudes of  $I_1$  and  $I_{2 \rightarrow 1}^*$ , (d) the gradient-enhanced image  $I_{12}^\dagger$ , (e) a detail of  $I_{12}^\dagger$ , (f) the gradient magnitude  $\Xi(M(I_{12}))$  of the stacked median image  $I_{12}$ . Best viewed in color. The reader is invited to zoom in on the electronic version of the manuscript in order to better appreciate the visual differences.



**Figure 5.** Application of the stacked median operator  $\boxplus$  for computing  $I_{12}$  from  $I_1 \cup I_{2 \rightarrow 1}^*$ . At pixel  $(x, y)$ , the stacked median operator takes the union of the corresponding  $3 \times 3$  local neighbourhoods for each image of the input set (in the example, the union of the red and green neighbourhoods, and the union of the orange and blue ones, missing data are represented in the figure as gray ticked boxes) and assigns its median intensity value to the point  $(x, y)$  in the new image. Best viewed in color.

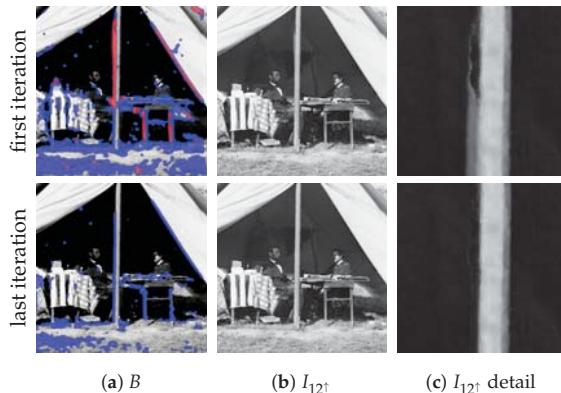
#### 2.4. Refinement

As already noted, the optical flow may be not perfect, causing the presence of wrong data in the image synthesis and hence in the data fusion process described in the previous step. To alleviate this issue, an iterative error-driven image correction step is introduced, where each iteration can be split into two sub-steps:

1. Detection. A binary correction mask is computed by considering the error image  $E = (I_1 - I_{12})^2$  the  $11 \times 11$  local window  $L(x, y)$  centered at each  $(x, y)$ . Given  $L'(x, y) \subseteq L(x, y)$  as the subset of pixels with intensity values lower than the 66% percentile on  $L(x, y)$ , the pixel  $(x, y)$  is marked as requiring adjustment if the square root of the average intensity value on  $L'(x, y)$  is higher than  $t = 16$  (chosen experimentally). This results in a binary correction mask  $B$  that is smoothed with a Gaussian kernel and then binarized again by a threshold value of 0.5. As clear from Figure 6a, using the percentile-based subset  $L'(x, y)$  is more robust than working with the whole window  $L(x, y)$ .

2. Adjustment. Data fusion is repeated again after updating pixels on  $I_{2 \rightarrow 1}^*$  that need to be adjusted with the corresponding ones of  $I_{12\uparrow}$ . Since  $I_{12\uparrow}$  is a sort of average between  $I_1$  and  $I_{2 \rightarrow 1}^*$ , the operation just described pushes marked pixels towards  $I_1$ . At the end of this step, the gradient enhanced image  $I_{12\uparrow}$  is also updated accordingly and, in case of no further iterations, it constitutes the final output.

Iterations stop when no more pixels to be adjusted are detected in the updated  $I_{12\uparrow}$  or when the maximum number of iterations is reached (see Figure 6). A maximum of four iterations is carried out, since it was verified experimentally that data fusion typically converges to  $I_1$  within this number of steps.



**Figure 6.** Refinement step (see Section 2.4). First (top row) and last (bottom row) iterations of the detection and adjustment sub-steps. (a) detection mask  $B$  at the beginning of the iteration, (b) updated  $I_{12\uparrow}$  at the end of the iteration and (c) details of (b). Pixels to be adjusted using  $L'$  ( $L$ ) are underlined in the images by saturating the red (blue) channel. By inspecting the details, it can be seen that the ghosting effect is removed. Best viewed in color. The reader is invited to zoom in on the electronic version of the manuscript in order to better appreciate the visual differences.

## 2.5. Guided Supersampling

Previous steps describe the original SMR implementation [23]. In order to preserve more fine details of the input images, a better image fusion is proposed hereafter, where the original coarse blended image  $I_{12}$  (Equation (4)) is employed to guide a refinement on the basis of supersampling (see Figure 7).

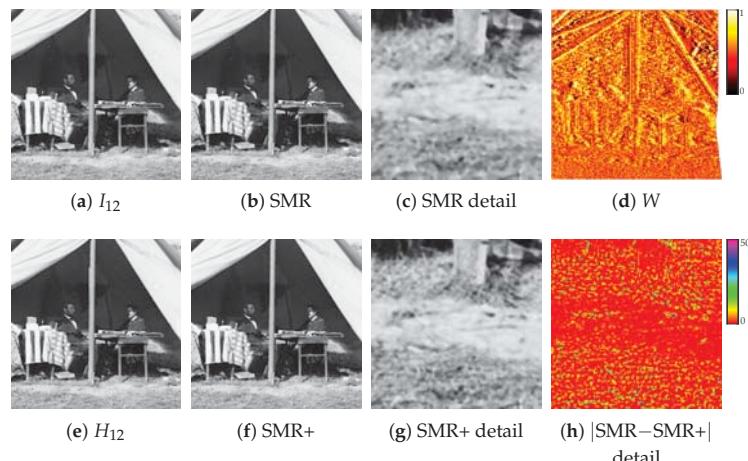
Let  $W_1$  denote the image obtained by averaging  $|I_1 - I_{12}|$  on a  $3 \times 3$  window, and similarly  $W_2$  the one obtained with  $|I_2 - I_{12}|$ . The weight mask  $W$  is computed as  $W_1 / (W_1 + W_2)$  pixelwise, followed by the convolution with a Gaussian with a standard deviation of four pixels (see Figure 7d). A value of  $W$  close to 0 (1) for a given pixel implies that the local neighborhood of that pixel in  $I_1$  ( $I_2$ ) is very likely less noisy and more artefact-free than  $I_2$  ( $I_1$ ). The mask  $W$  is used to define a *weighted stacked median*

$$H_{12} = \square_W(I_1^{\times 2}, I_{2 \rightarrow 1}^{*\times 2}) \quad (11)$$

where the superscript  $\times 2$  indicates the bicubic rescaling by a factor two for supersampling (see Figure 7e). Explicitly, the weighted stacked median at  $(x, y)$  is obtained as the median of the intensities of  $V_1(x, y) \cup V_2(x, y)$ , where  $V_1(x, y) \subseteq I_1$  is the subset of the pixels in the  $3 \times 3$  local neighbourhood of  $(x, y)$  containing the  $[w \times \min(1 - W(x, y), w')]$  intensity values closest to  $I_{12}^{\times 2}(x, y)$ , with  $w = 3^2 \times 2$  and  $w' = (3^2 + 1) / (2 \times 3^2 + 1)$ , and likewise for  $V_2(x, y) \subseteq I_2$  containing the pixels with the  $[w \times \min(W(x, y), w')]$  closest values. In other words, the number of considered samples for the median taken from each image

is proportional to the weight  $W(x, y)$ . The cardinalities of the subsets  $V_1$  and  $V_2$  for the different weight ranges are explicitly shown in Table 1.

The high resolution blended image  $H_{12}$  replaces  $I_{12}$  in the next steps of the method (see Sections 2.3 and 2.4),  $I_1$  and  $I_2$  being also replaced accordingly by  $I_1^{\times 2}$  and  $I_2^{\times 2}$ . The final output is scaled down to the original input size. With respect to the original SMR implementation, the use of guided supersampling in SMR+ preserves better fine details, also improving further the restoration process (compare Figure 7c,g). Notice that, after each refinement sub-step (see Section 2.4), the coarse  $I_{12}$  image needed to guide the process is generated by the stacked median between  $I_1$  and  $I_{12}^{\times 2}$  scaled down to the original size.



**Figure 7.** Guided supersampling step (see Section 2.5): (a) SMR stacked median  $I_{12}$ , (b) final restored image and (c) details of it, (d) weight mask  $W$  for the guided supersampling, (e) SMR+ weighted stacked median  $H_{12}$ , (f) final restored image, (g) a detail of it, and (h) its differences with respect to the SMR output. Best viewed in color. The reader is invited to zoom in on the electronic version of the manuscript in order to better appreciate the visual differences.

**Table 1.** The cardinality of the sets  $V_1(x, y)$  and  $V_2(x, y)$  according to the weight  $W(x, y)$  range (see Section 2.5).

	inf $W(x, y)$	0.00	0.05	0.11	0.16	0.21	0.26	0.32	0.37	0.42	0.47	0.53	0.58	0.63	0.68	0.74	0.79	0.84	0.89	0.95	0.95
	sup $W(x, y)$	0.05	0.11	0.16	0.21	0.26	0.32	0.37	0.42	0.47	0.53	0.58	0.63	0.68	0.74	0.79	0.84	0.89	0.95	1.00	
$ V_1(x, y) $	9	9	9	9	9	9	9	9	9	9	9	8	7	6	5	4	3	2	1	0	
$ V_2(x, y) $	0	1	2	3	4	5	6	7	8	9	9	9	9	9	9	9	9	9	9	9	

### 3. Evaluation

#### 3.1. Dataset

In order to evaluate the proposed approach, we built a new dataset including historical stereo pairs from different sources. The left frames of the selected stereo pairs are shown as reference in Figure 8.

A first set of seven stereo pairs belongs to the collection of stereograms by Anton Hautmann, one of the most active photographers in Florence between 1858 and 1862. Part of Hautmann's collection is described in [21]. The seven stereo pairs used in this work depict different viewpoints of Piazza Santissima Annunziata in Florence as it was in the middle of the 19th century. Inspecting these photos (see Figure 8, red frames), it can be noticed that the image quality is very poor. In particular, the pairs are quite noisy, with low definition and contrast, include saturated or blurred areas and also show scratches and stains.



**Figure 8.** Left frames for some historical stereo pairs. Image frames for Hautmann’s, Stereoscopic Photos and USGS datasets are framed, respectively, in red, blue and green. Best viewed in color and zoomed in with the electronic version of the manuscript.

A second set includes 35 stereo pairs and increases the original set of ten images employed in [23]. These stereo pairs have been gathered from the Stereoscopic History Instagram account (<https://www.instagram.com/stereoscopichistory/>, accessed on 1 April 2021, see Figure 8, blue frames for some examples) and contain landscape pictures of urban and natural scenes as well as individual or group portraits. This set is the most challenging one, since its images are heavily corrupted by noise and other artefacts.

A third set of five images was collected from the U.S. Geological Survey (USGS) Historical Stereoscopic Photos account on Flickr (<https://www.flickr.com/photos/usgeologicalsurvey/>, accessed on 1 April 2021), and represents natural landscapes (see Figure 8, green frames), except for the last one which also includes two horsemen with their mounts. The quality of these images is similar to that of the first set, but strong vignetting effects are also present.

### 3.2. Compared Methods

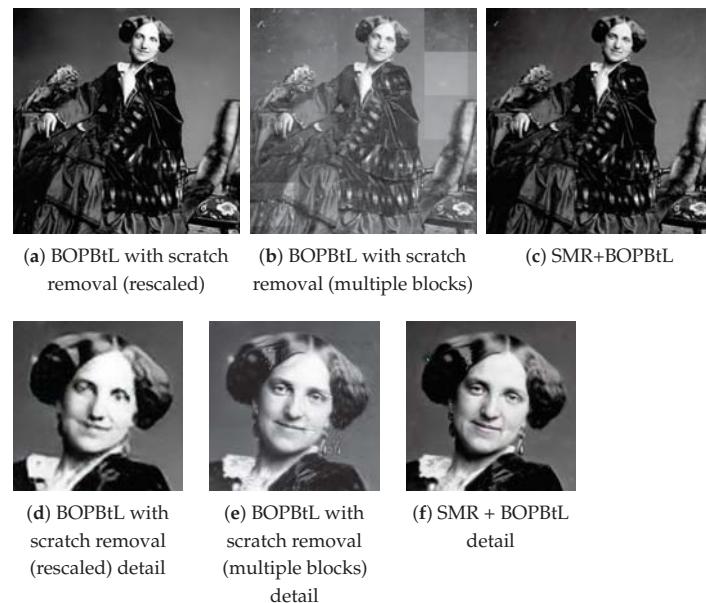
The proposed SMR and SMR+ are compared against Block Matching 3D (BM3D) [7], Deep Image Prior (DIP) [13] and the recent BOPBtL [26]. BM3D and DIP are, respectively, a handcrafted and deep generic denoising methods, while BOPBtL is a deep network specifically designed for old photo restoration. These methods currently represent the state-of-the-art in this research area.

For BM3D, the legacy version was employed, since, according to our preliminary experiments, the new version including correlated noise suppression did not work well for our kind of images. The BM3D  $\sigma$  parameter, the only one present, was set to 7 and 14, values that, according to our experiments, gave the best visual results. In particular,  $\sigma = 14$  seems to work better than  $\sigma = 7$  in the case of higher resolution images. Besides applying the standard BM3D on the reference image, a modified version of this method was implemented in order for BM3D to benefit from the stereo auxiliary data. Since BM3D

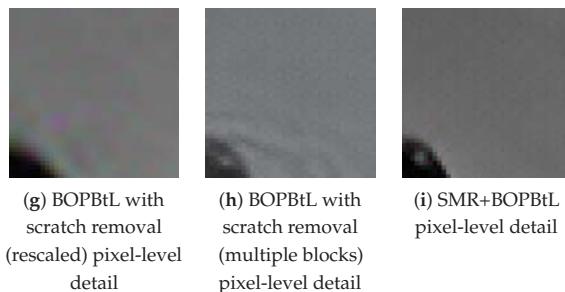
exploits image self-correlation to suppress noise, the modified BM3D generates auxiliary sub-images by placing side by side two corresponding  $96 \times 96$  patches from  $I_1$  and  $I_{2 \rightarrow 1}^*$ , then runs the original BM3D on each sub-image and finally generates the output by collecting the blocks from each sub-image corresponding to the  $32 \times 32$  central  $I_1$  patches. No difference in the results with respect to the standard BM3D was observed, which plausibly implies that corresponding patches for  $I_1$  and  $I_{2 \rightarrow 1}^*$  are not judged as similar to each other by BM3D.

In the case of DIP, the borders of the input images were cropped due to network architectural constraints: These missing parts were replaced with content from the original input images.

Concerning BOPBtL, the scratch removal option was disabled since it caused the network to crash. This is a known issue related to the high memory requirement exceeding the standard 12 GB GPU amount to run the network on standard image input (<https://github.com/microsoft/Bringing-Old-Photos-Back-to-Life/issues/>, accessed on 1 April 2021), and does not occur only when the input image size is small. To circumvent this problem, two solutions were attempted, yet without satisfying results. Specifically, in the first solution, the input image was rescaled to a fixed size (from 50% to 33% of the its original size), but the final result lost too many details (see Figure 9a). In the second solution, the input was processed in separated blocks, causing a lack of global consistency in the output (see Figure 9b). Moreover, in both solutions, the chessboard artefact effect, typical of many deep networks that resynthesize images, looked more evident than in the original BOPBtL implementation. BOPBtL was employed to post-process the output of SMR+, which was denoted as SMR+BOPBtL in the results (see Figure 9c).



**Figure 9. Cont.**

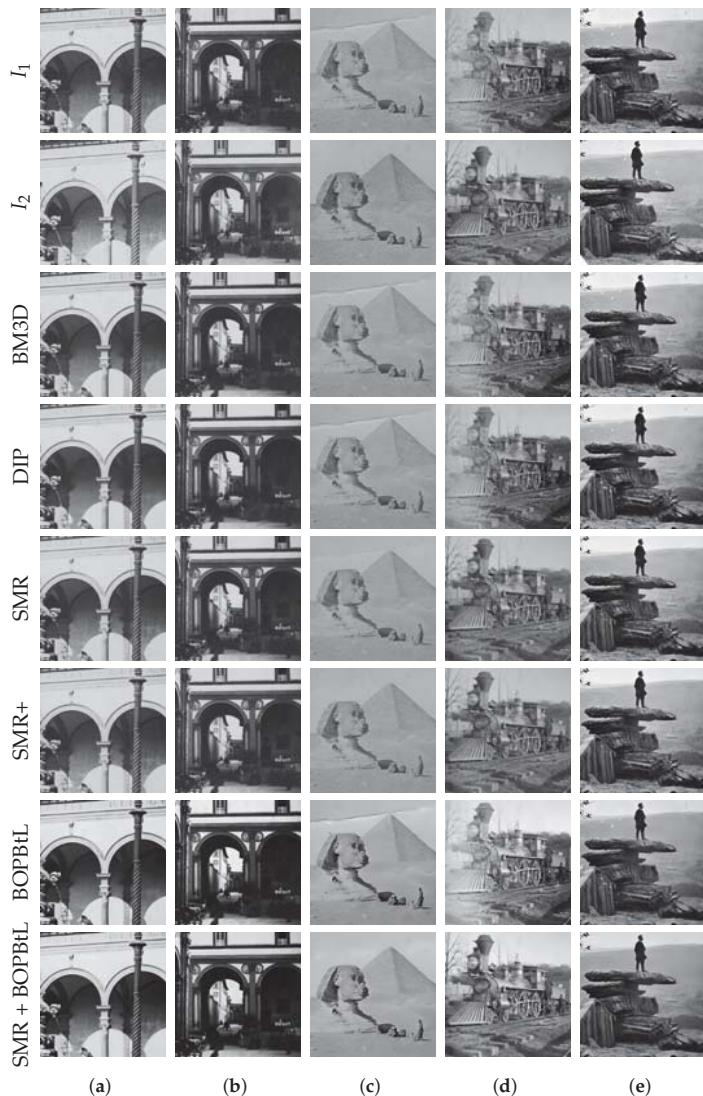


**Figure 9.** Results of BOPBtL with scratch removal or in combination with SMR+ on the same stereo pair of Figure 1. Notice that the visual pleasant results of (a) are due to the frequency cutoff caused by rescaling and disappear at a larger viewing scale such as (d). Best viewed in color. The reader is invited to zoom in on the electronic version of the manuscript in order to better appreciate the visual differences.

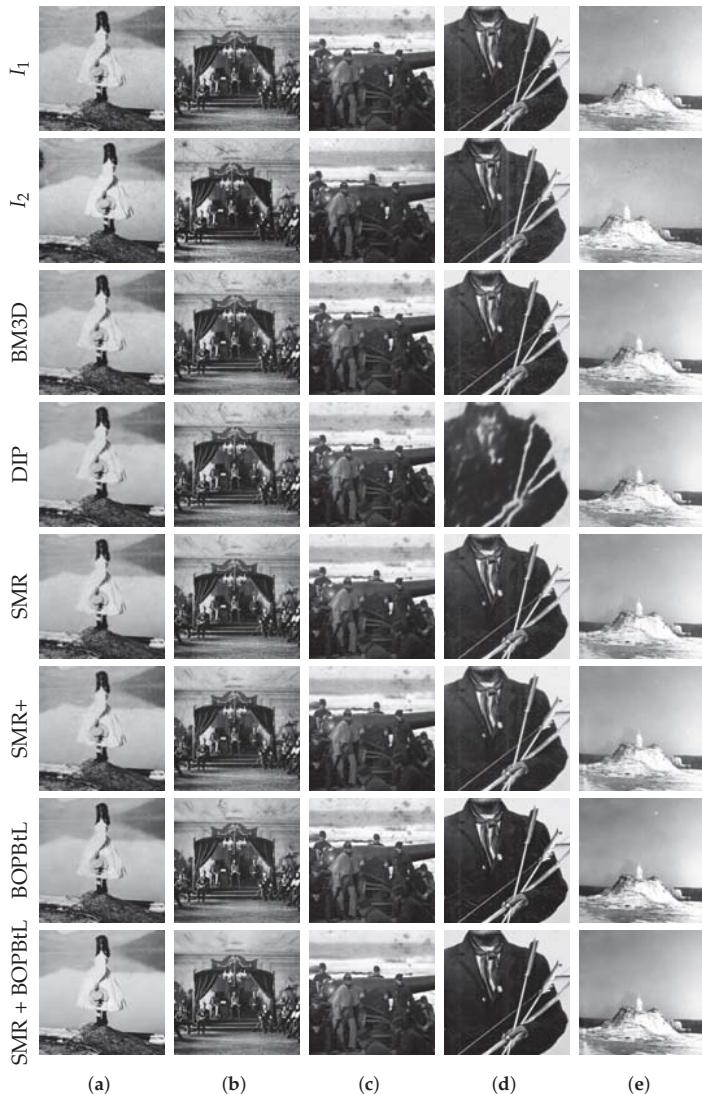
### 3.3. Results

Figures 10 and 11 show some examples of the results obtained with the compared methods. For a thorough visual qualitative evaluation, the reader is invited to inspect the full-resolution results obtained on the whole dataset, which are included in the additional material (<https://drive.google.com/drive/folders/1Fmsm50bMMDSd0z4jXOhCZ3hPDIXdwMUL>). From a direct visual inspection of the results, BM3D and DIP often seem to oversmooth relevant details in the image, with BM3D producing somewhat better results than DIP, which sometimes simply fails to obtain a reasonable output (see Figure 11d, row DIP). BOPBtL is able to bring out fine details, providing altogether a locally adaptive smoothing and contrast enhancement of the image, with satisfactory results. Nevertheless, none of the previous methods is able to detect and compensate for dust, scratches and other kinds of artefacts that conversely may even be amplified in the restoration process, as one can check by locating dust spots and sketches in Figure 10e, rows BM3D, DIP and BOPBtL. This problem is mostly evident for BOPBtL, where image artefacts are heavily boosted together with finer details.

Conversely, SMR-based methods are able to solve these issues by exploiting the additional information present in the auxiliary image, with the exception of very severe conditions such as the stains appearing in the right skyline of Figure 11c, for which, anyways, SMR-based methods still get the best restoration of all. SMR-based methods also successfully enhance the image contrast, as it happens for the window in the dark spot under the right arcade in Figure 10b, rows SMR and SMR+. When image degradations are even more severe than that, good results can nevertheless be obtained by forcing the illumination of the auxiliary image into the reference (see Section 2.2), as done for Figure 10d, rows SMR, SMR+ and SMR+BOPBtL. Concerning the guided supersampling introduced for SMR+, this is able not only to preserve high frequency details (see again Figure 7), but also to better clean-up the image, as one can notice by inspecting the removed scratch from Figure 10c, row SMR+. Guided supersampling also alleviates spurious artefacts arising from inaccurate optical flow estimation as in the case of the light pole of Figure 10a (compare rows SMR and SMR+). Only in few cases of very inaccurate optical flow estimation is SMR+ unable to fix inconsistencies and generates some spurious artefacts as in the bottom-left white scratch in Figure 11e, rows SMR+ and SMR + BOPBtL. Finally, it can be noted that SMR + BOPBtL is able to take the best from both the methods, i.e., the artefact removal from SMR+ and the image enhancement from BOPBtL, and provides very visually striking results.



**Figure 10.** Qualitative visual comparison of the methods under test. Best viewed in color. The reader is invited to zoom in on the electronic version of the manuscript in order to better appreciate the differences.



**Figure 11.** Qualitative visual comparison of the methods under test. Best viewed in color. The reader is invited to zoom in on the electronic version of the manuscript in order to better appreciate the differences.

Table 2 reports the score obtained by the compared methods on the images discussed so far according to popular no-reference quality assessment metrics. Specifically, scores are reported for the Blind / Referenceless Image Spatial Quality Evaluator (BRISQUE) [29], the Naturalness Image Quality Evaluator (NIQE) [30] and Perception based Image Quality Evaluator (PIQE) [31]. Due to the lack of ground-truth clean data and of a well-defined image model for the generation of synthetic images with the same characteristics of the input image under evaluation, image quality measurements requiring a reference image such as the Structural Similarity Index (SSIM) [32] cannot be applied. By inspection of the scores obtained, it clearly emerges that these quality metrics do not reflect the human visual

judgment, hence they are unsuitable for a reliable quantitative evaluation in this specific application scenario. In particular, there is no agreement among the various metrics and, in about half of the cases, the input image even gets a better score than the restored one, in contrast with the human visual assessment. Furthermore, SMR+ and SMR + BOPBtL obtain worse scores than the original images or BOPBtL in the cases where SMR-methods successfully cleaned the image by removing strong image artefacts, again in contrast with human judgment (see Figure 11b,d). A possible explanation of this behavior is that these metrics only rely on low-level, local image properties and not on high-level, semantic image characteristics. Hence, they are unable to distinguish between fine image details and artefacts. Nevertheless, according to Table 2, SMR+, with or without BOPBtL, shows good results under these blind quality assessment metrics, implying that it is able not only to remove structural artefacts from the original image, but also to maintain high quality visual details besides the semantic interpretation of the scene.

**Table 2.** No-reference assessment metric results (lower values are better). Values in bold indicate the best score among the compared methods. Scores that are better in the original images than in the restored ones are underlined.

		$I_1$	BM3D	DIP	SMR	SMR+	BOPBtL	SMR+ BOPBtL
Figures 1 and 9	BRISQUE	41.89	54.34	51.47	53.11	43.46	<b>24.15</b>	24.20
	NIQE	4.23	5.31	5.31	5.09	<b>3.98</b>	4.09	4.24
	PIQE	45.97	<u>78.93</u>	85.33	50.60	46.35	<b>22.55</b>	25.90
Figure 10a	BRISQUE	<u>10.74</u>	46.03	31.11	42.18	33.06	<b>25.41</b>	31.37
	NIQE	<u>2.79</u>	3.83	3.94	<b>3.28</b>	3.76	4.05	4.08
	PIQE	<u>25.02</u>	79.24	81.50	43.32	<b>28.09</b>	38.50	35.35
Figure 10b	BRISQUE	<u>9.84</u>	48.68	35.95	41.57	29.69	<b>14.17</b>	34.69
	NIQE	3.16	4.07	3.92	<b>2.92</b>	3.34	3.65	4.01
	PIQE	29.73	<u>78.53</u>	78.16	37.26	<b>23.61</b>	29.98	34.31
Figure 10c	BRISQUE	<u>9.26</u>	44.97	31.28	38.29	33.94	<b>12.13</b>	19.06
	NIQE	<u>2.79</u>	4.22	4.11	<b>3.47</b>	4.04	5.43	5.31
	PIQE	<u>15.80</u>	60.33	53.28	42.81	23.02	20.30	<b>20.00</b>
Figure 10d	BRISQUE	14.57	31.93	22.82	36.91	25.66	15.89	<b>10.96</b>
	NIQE	<u>2.61</u>	<b>3.11</b>	3.72	3.49	3.65	3.97	3.62
	PIQE	<u>9.31</u>	43.23	52.66	38.28	24.24	<b>10.48</b>	11.76
Figure 10e	BRISQUE	<u>12.85</u>	30.58	28.31	31.95	<b>22.40</b>	29.13	28.87
	NIQE	<u>2.17</u>	2.26	3.30	3.13	<b>2.92</b>	4.05	3.97
	PIQE	27.52	42.54	45.40	40.00	24.43	<b>14.67</b>	16.92
Figure 11a	BRISQUE	42.58	48.03	40.26	51.88	41.23	38.48	<b>39.21</b>
	NIQE	<u>3.80</u>	4.77	4.97	4.66	<b>3.93</b>	4.57	4.75
	PIQE	26.39	74.37	79.44	45.89	36.91	<b>13.28</b>	14.60
Figure 11b	BRISQUE	39.15	49.22	53.80	45.41	40.85	<b>14.75</b>	17.74
	NIQE	4.33	5.43	5.78	4.93	<b>4.15</b>	4.32	4.56
	PIQE	28.96	82.41	84.95	46.49	38.68	<b>15.54</b>	17.70
Figure 11c	BRISQUE	30.43	52.90	55.07	52.86	39.59	25.54	<b>20.06</b>
	NIQE	<u>3.13</u>	5.22	5.53	4.25	<b>3.20</b>	4.59	4.36
	PIQE	<u>17.20</u>	85.95	88.53	43.98	30.33	<b>25.39</b>	27.83
Figure 11d	BRISQUE	28.40	45.63	47.19	41.24	31.51	<b>22.09</b>	23.47
	NIQE	2.11	4.17	6.28	3.89	<b>2.85</b>	3.49	3.85
	PIQE	31.65	72.88	94.84	48.02	36.64	<b>20.68</b>	22.81
Figure 11e	BRISQUE	40.12	38.54	37.95	<b>20.01</b>	22.15	38.12	22.07
	NIQE	6.27	3.49	4.08	2.84	<b>3.06</b>	4.60	4.42
	PIQE	58.45	51.79	48.00	19.77	<b>13.28</b>	13.35	11.45

Concerning running times, BM3D, BOPBtL and DIP require respectively 10 s, 30 s and 20 min on average for processing the dataset images, while SMR and SMR+ take respectively 6 min and 9 min. The running environment is a Ubuntu 20.04 system running on an Intel Core i7-3770K with 8 GB of RAM equipped with a 12 GB NVIDIA Titan XP. BM3D is a Matlab optimized .mex file, BOPBtL and DIP implementations run on Pytorch exploiting GPU acceleration, while, with the exception of RAFT optical flow estimation, SMR and SMR+ are based on non-optimized Matlab code running on CPU. For both SMR and SMR+ the times include the image resynthesis and color correction steps that take 4.5 min altogether on average. Under these considerations, both SMR and SMR+ running times are reasonable for offline applications. None of the compared methods can be used for real-time applications, as in the best case corresponding to BM3D, 10 s are required for processing the input image.

#### 4. Conclusions and Future Work

This paper proposed a novel method for the fully automatic restoration of historical stereo photographs. By exploiting optical flow, the auxiliary view of the stereo frame is geometrically and photometrically registered onto the reference view. Restoration is then carried out by fusing the data from both images according to our stacked median approach followed by gradient adjustments aimed at preserving details. Guided supersampling is also introduced and successfully applied for enhancing finer details and simultaneously providing a more effective artefact removal. Finally, an iterative refinement step driven by a visual consistency check is performed in order to remove the artefacts due to optical flow estimation errors in the initial phase.

Results on several historical stereo pairs show the effectiveness of the proposed approach that is able to remove most of the image defects including dust and scratches, without excessive smoothing of the image content. The approach works better than its single-image denoising competitors, thanks to the ability of exploiting stereo information. As a matter of fact, single-image methods have severe limitations in handling damaged areas, and usually produce more blurry results. Nevertheless, experimental results show that single image BOPBtL, when cascaded with our approach into SMR + BOPBtL, can achieve remarkably good performances.

Future work will investigate novel solutions to refine the optical flow in order to reduce pixel mismatches. A further research direction will be towards consolidating the stacked median approach as an image blending technique. Finally, the proposed method will be extended and adapted to the digital restoration of historical films.

**Author Contributions:** Conceptualization, F.B.; methodology, F.B. and M.F.; software, F.B and M.F.; validation, F.B., M.F. and C.C.; formal analysis, F.B., M.F. and C.C.; investigation, F.B. and M.F.; resources, F.B., M.F. and C.C.; data curation, F.B., M.F. and C.C.; writing—original draft preparation, F.B.; writing—review and editing, M.F., C.C. and F.B.; visualization, F.B., M.F. and C.C.; supervision, F.B. and C.C.; project administration, F.B. and C.C.; funding acquisition, F.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Italian Ministry of University and Research (MUR) under the program PON Ricerca e Innovazione 2014–2020, cofunded by the European Social Fund (ESF), CUP B74I18000220006, id. proposta AIM 1875400, linea di attività 2, Area Cultural Heritage.

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable

**Data Availability Statement:** Additional material including code, dataset and evaluation results are freely available online at <https://drive.google.com/drive/folders/1Fmsm50bMMDSD0z4jXOhCZ3hPDXdwMUL>.

**Acknowledgments:** The Titan Xp used for this research was generously donated by the NVIDIA Corporation. We would like to thank Costanza Caraffa and Ute Dercks at Photothek des Kunsthis-

torischen Instituts in Florenz–Max-Planck-Institut for allowing the reproduction of the photos in this paper. Hautmann’s collection digital scans: ©Stefano Fancelli/KHI.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

1. Ardizzone, E.; De Polo, A.; Dindo, H.; Mazzola, G.; Nanni, C. A Dual Taxonomy for Defects in Digitized Historical Photos. In Proceedings of the 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 26–29 July 2009; pp. 1166–1170.
2. Kokaram, A.C. *Motion Picture Restoration: Digital Algorithms for Artefact Suppression in Degraded Motion Picture Film and Video*; Springer: Berlin/Heidelberg, Germany, 1998.
3. Tegolo, D.; Isgrò, F. A genetic algorithm for scratch removal in static images. In Proceedings of the International Conference on Image Analysis and Processing (ICIAP2001), Palermo, Italy, 26–28 September 2001; pp. 507–511.
4. Stanco, F.; Tenze, L.; Ramponi, G. Virtual restoration of vintage photographic prints affected by foxing and water blots. *J. Electron. Imaging* **2005**, *14*, 043008. [[CrossRef](#)]
5. Besserer, B.; Thiré, C. Detection and Tracking Scheme for Line Scratch Removal in an Image Sequence. In Proceedings of the European Conference on Computer Vision (ECCV2004), Prague, Czech Republic, 11–14 May 2004; pp. 264–275.
6. Criminisi, A.; Perez, P.; Toyama, K. Object removal by exemplar-based inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2003), Madison, WI, USA, 16–22 June 2003; Volume 2.
7. Dabov, K.; Foi, A.; Katkovnik, V.; Egiazarian, K. Image Denoising by Sparse 3D Transform-Domain Collaborative Filtering. *IEEE Trans. Image Process.* **2007**, *16*, 2080–2095. [[CrossRef](#)] [[PubMed](#)]
8. Chen, F.; Zhang, L.; Yu, H. External Patch Prior Guided Internal Clustering for Image Denoising. In Proceedings of the IEEE International Conference on Computer Vision (ICCV2015), Santiago, Chile, 7–13 December 2015; pp. 603–611.
9. Buades, A.; Lisani, J.; Miladinović, M. Patch-Based Video Denoising With Optical Flow Estimation. *IEEE Trans. Image Process.* **2016**, *25*, 2573–2586. [[CrossRef](#)]
10. Nasrollahi, K.; Moeslund, T.B. Super-resolution: A comprehensive survey. *Mach. Vis. Appl.* **2014**, *25*, 1423–1468. [[CrossRef](#)]
11. Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. Image Super-Resolution Via Sparse Representation. *IEEE Trans. Image Process.* **2010**, *19*, 2861–2873. [[CrossRef](#)] [[PubMed](#)]
12. Castellano, G.; Vessio, G. Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. *Neural Comput. Appl.* **2021**. [[CrossRef](#)]
13. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Deep Image Prior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2018), Salt Lake City, UT, USA, 18–23 June 2018.
14. Wang, Z.; Chen, J.; Hoi, S.C.H. Deep Learning for Image Super-resolution: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [[CrossRef](#)] [[PubMed](#)]
15. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual Dense Network for Image Restoration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 2480–2495. [[CrossRef](#)]
16. Teed, Z.; Deng, J. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In Proceedings of the European Conference on Computer Vision (ECCV2020), Glasgow, UK, 23–28 August 2020.
17. Jeon, D.S.; Baek, S.; Choi, I.; Kim, M.H. Enhancing the Spatial Resolution of Stereo Images Using a Parallax Prior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2018), Salt Lake City, UT, USA, 18–23 June 2018; pp. 1721–1730.
18. Zhou, S.; Zhang, J.; Zuo, W.; Xie, H.; Pan, J.; Ren, J.S. DAVANet: Stereo Deblurring With View Aggregation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2019), Long Beach, CA, USA, 16–20 June 2019; pp. 10988–10997.
19. Yan, B.; Ma, C.; Bare, B.; Tan, W.; Hoi, S. Disparity-Aware Domain Adaptation in Stereo Image Restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2020), Seattle, WA, USA, 14–18 June 2020; pp. 13176–13184.
20. Schindler, G.; Dellaert, F. 4D Cities: Analyzing, Visualizing, and Interacting with Historical Urban Photo Collections. *J. Multimed.* **2012**, *7*, 124–131. [[CrossRef](#)]
21. Fanfani, M.; Bellavia, F.; Bassetti, G.; Argenti, F.; Colombo, C. 3D Map Computation from Historical Stereo Photographs of Florence. *IOP Conf. Ser. Mater. Sci. Eng.* **2018**, *364*, 012044. [[CrossRef](#)]
22. Luo, X.; Kong, Y.; Lawrence, J.; Martin-Brualla, R.; Seitz, S. KeystoneDepth: Visualizing History in 3D. *arXiv* **2019**, arXiv:1908.07732.
23. Fanfani, M.; Colombo, C.; Bellavia, F. Restoration and Enhancement of Historical Stereo Photos through Optical Flow. In Proceedings of the ICPR Workshop on Fine Art Pattern Extraction and Recognition (FAPER), Milan, Italy, 18 September 2021.
24. McCann, J.J.; Rizzi, A. *The Art and Science of HDR Imaging*; John Wiley & Sons Inc.: Chichester, UK, 2011.
25. Sherrod, A. *Game Graphic Programming*; Course Technology: Boston, MA, USA, 2008.
26. Wan, Z.; Zhang, B.; Chen, D.; Zhang, P.; Chen, D.; Liao, J.; Wen, F. Bringing Old Photos Back to Life. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–18 June 2020.

27. Chambon, S.; Crouzil, A. Similarity measures for image matching despite occlusions in stereo vision. *Pattern Recognit.* **2011**, *44*, 2063–2075. [[CrossRef](#)]
28. Bellavia, F.; Colombo, C. Dissecting and Reassembling Color Correction Algorithms for Image Stitching. *IEEE Trans. Image Process.* **2018**, *27*, 735–748. [[CrossRef](#)]
29. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. [[CrossRef](#)] [[PubMed](#)]
30. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a “Completely Blind” Image Quality Analyzer. *IEEE Signal Process. Lett.* **2013**, *20*, 209–212. [[CrossRef](#)]
31. Venkatanath, N.; Praneeth, D.; Chandrasekhar, B.; Channappayya, S.; Medasani, S. Blind image quality evaluation using perception based features. In Proceedings of the 21st National Conference on Communications (NCC), Mumbai, India, 27 February–1 March 2015; pp. 1–6.
32. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]



## Article

# Analysis of Diagnostic Images of Artworks and Feature Extraction: Design of a Methodology

Annamaria Amura <sup>1,\*</sup>, Alessandro Aldini <sup>1</sup>, Stefano Pagnotta <sup>2</sup>, Emanuele Salerno <sup>3</sup>, Anna Tonazzini <sup>3</sup> and Paolo Triolo <sup>4</sup>

- <sup>1</sup> Department of Pure and Applied Sciences, University of Urbino Carlo Bo, 61029 Urbino, Italy; alessandro.aldini@uniurb.it  
<sup>2</sup> Department of Earth Sciences, University of Pisa, 56126 Pisa, Italy; stefano.pagnotta@unipi.it  
<sup>3</sup> National Research Council of Italy, Institute of Information Science and Technologies, 56124 Pisa, Italy; emanuele.salerno@isti.cnr.it (E.S.); anna.tonazzini@isti.cnr.it (A.T.)  
<sup>4</sup> Department of Earth Sciences of the Environment and Life, Methodologies for Conservation and Restoration of Cultural Heritage, University of Genoa, 16126 Genova, Italy; triolox@libero.it  
\* Correspondence: annamaria.amura@uniurb.it

**Abstract:** Digital images represent the primary tool for diagnostics and documentation of the state of preservation of artifacts. Today the interpretive filters that allow one to characterize information and communicate it are extremely subjective. Our research goal is to study a quantitative analysis methodology to facilitate and semi-automate the recognition and polygonization of areas corresponding to the characteristics searched. To this end, several algorithms have been tested that allow for separating the characteristics and creating binary masks to be statistically analyzed and polygonized. Since our methodology aims to offer a conservator-restorer model to obtain useful graphic documentation in a short time that is usable for design and statistical purposes, this process has been implemented in a single Geographic Information Systems (GIS) application.

**Keywords:** cultural heritage; diagnostic images; image analysis; feature extraction; documentation; geographic information systems (GIS)

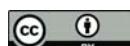


**Citation:** Amura, A.; Aldini, A.; Pagnotta, S.; Salerno, E.; Tonazzini, A.; Triolo, P. Analysis of Diagnostic Images of Artworks and Feature Extraction: Design of a Methodology. *J. Imaging* **2021**, *7*, 53. <https://doi.org/10.3390/jimaging7030053>

Academic Editor:  
Giovanna Castellano

Received: 8 February 2021  
Accepted: 8 March 2021  
Published: 12 March 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The artworks undergo profound changes in time due to several factors: the natural aging of materials, pathologies of degradation, wrong restoration or remaking that introduce new materials and chemical interactions. For this reason, any technique for detecting and reporting what is not directly visible or perceptible is an essential means of diagnostic investigation. This need is currently widely met by techniques inspired by remote sensing, such as multispectral and hyperspectral imaging, as they can provide information on the composition of materials without taking samples. The technical information related to the nature and conditions of the artwork is transcribed in the graphic documentation.

The documentation refers to the systematic collection of information derived from the diagnostic investigation, restoration, monitoring, and maintenance performed on cultural heritage. Specifically, the graphic documentation, also called the thematic map, is the primary tool for communication and synthesis of the information collected on the nature and conditions of the artwork, which are transcribed into geometrically correct drawings and translated into conventional graphic symbols [1,2]. Thematic maps are generally used by different types of professionals operating at different times and in different ways, representing the formal and unequivocal means of communication, comparison, and guidance for subsequent conservation and preservation operations. The graphic documentation should always be well archived, accessible, and usable; therefore, it should be possible to obtain or arrive at the information when and where it is needed.

In the documentation process, these graphic drawings are intended for critical analysis of data. They differ from the artifact's photographic reproduction, which detects and

reproduces all its complexities in an undifferentiated way, without a critical/interpretative filter. The vector drawing allows us to realize a process of synthesis, discrepancy, and characterization of data, making the results immediately readable and statistically analyzable. Although software for the documentation of three-dimensional models is being developed very slowly and pioneeringly, graphic documentation in the form of thematic maps is always required during a conservation or preservation intervention.

Both for artifacts with greater three-dimensionality and for artifacts with reduced three-dimensionality, the vector graphic drawing is always based on a two-dimensional photographic reproduction of the artifact, which is often not geometrically correct. The artifact is then photographed in all its sides at 360° and a thematic map is created for each side or prospect and the graphic documentation operation is performed during the entire intervention. In the current practice, this process is carried out through manual drawing by restorers, so it is strongly influenced by their skills.

In these subjective analysis processes, the operations of area graphicization and interpretation of the characteristics constitute a joint phase. Indeed, those who perform the mapping outline the areas of interest directly following the boundaries dictated by their experience and visual perception.

To date, the automatic extraction of drawings from raster images has only been made in archaeology by a specific technique called Stippling. It has been developed to produce illustrations in raster format, extracted from photographs of archaeological objects [3]. Unfortunately, these techniques do not meet the requirements of graphic documentation in restoration.

Our research aims to study a quantitative analysis methodology to facilitate and semi-automate the recognition and vectorization of areas corresponding to the characteristics under consideration. To this aim, some segmentation algorithms have been tested to separate the characteristics and facilitate identifying the areas and their vectorization. The choice to go beyond the analysis carried out directly on the pixel areas by performing a vectorization was dictated by the fact that the documentation in the restoration of any artifact requires a series of vector drawings, non-illustrative and non-raster, without nuances and with closed polygons topographically consistent with each other. So, the research topic we are developing is not only about one algorithm, but is about a formal methodology involving the cascaded application of a series of algorithms, which has been consolidated over the years. Our work's novelty is the methodology itself rather than the algorithms used to apply it in practice. This contribution includes the description of the supporting algorithms and the software tools implementing them. The study of it would allow for the full reproducibility of the methodology in practice.

This methodology has been developed in three years, during which it has been tested on several paintings on canvas, mosaics, frescoes/wall paintings, and paper/parchment artifacts, involving the profiles of diagnosticians, art historians, conservator/restorers and Geographic Information Systems (GIS) professionals. The model presented in this document is the one that has allowed us to obtain the best results in all the tested case studies.

The rest of the paper is organized as follows. In Section 2, the main problems present today in the documentation process are analyzed, and research projects focused on their solution are mentioned. We also mention the graphical documentation software, highlighting those tools that can support our methodology's semi-automatic implementation, such as, e.g., Geographic Information Systems (GIS).

In Section 3, we present the various stages of our methodology, which, in Section 4, is applied to a canvas painting. Finally, in Section 5, we conclude the paper and discuss potential future work.

## 2. State of the Art

Although the importance of documentation has been widely recognized and considerable experience has been gained in applying innovative documentation systems [4], there are still many unsolved problems in analyzing and digitizing artifacts.

## 2.1. Creation of Thematic Maps

Standardized techniques of architectural and archaeological survey (especially those related to the detection of the constituent materials of the external surfaces of buildings) have been the guide for the development of current thematic maps for the planning of conservation, as well as buildings and monuments of all other categories of cultural heritage. For this reason, only in the context of historical architectural monuments and archaeological sites can restorers rely on professionals, e.g., architects, and ad hoc standards for the generation of thematic maps. For interior decorations of buildings and small movable objects such as painted canvases and wood panels, sculptures, and utensils, the conservator-restorer tries to conform to the same architectural survey criteria, often without following a standard methodology.

The lack of standardization concerns four fundamental aspects: the modalities of photographic acquisition, the modalities of post-production and study of diagnostic investigations, the textual/graphic vocabulary of thematic maps and the use of software to create it. In this section, we focus on using software to create thematic maps and on the study and post-production of diagnostic investigations, because the other two aspects involve complex issues, often dependent on the type of artifact and the regulations in force in each country.

However, we think it is useful to provide some clarification about these issues: the modalities of photographic acquisition change according to the typology of the artifacts and the techniques used to respect the geometric correctness of the artifact represent a very broad field of research. In Italy, the only official document has been drawn up by the Central Institute for Catalogue and Documentation (ICCD) in Rome [5]. Concerning the textual and graphic vocabulary, very few standards are currently in use for digital architectural design [6] and the UNI Beni Culturali NorMal (Norme Materiale Lapideo) standards [1] mainly refer to undecorated stone material, mortars, and ceramics.

## 2.2. Software in Use

Currently, the most used software and platforms supporting graphic documentation are Computer-aided design (CAD) [7,8] and Geographic Information System (GIS) [9]. An interesting approach is offered in the geographical setting since the problems related to statistical analysis of satellite images and cartography creation are similar to cultural heritage documentation. GIS technologies offer flexible image analysis and data management toolboxes by integrating various functionalities, data types, and formats. In the field of restoration, implementations of GIS were developed in Italy in the 1990s through a project called “Carta del rischio” [10]. GIS and CAD functionalities are also merged into hybrid platforms, like SICAR® [11] a web-based geographic information system officially adopted in 2012 by the Italian Ministry of Cultural Heritage and Activities and Tourism.

All these tools suffer from some significant practical limitations. CAD is optimal for the mathematical processing of geometric data. However, its use is particularly challenging when the graphic survey to be produced is characterized by irregular and highly jagged edges and shapes. Using CAD drawing tools, operators tend to approximate the area's perimeters making the edges inaccurate. Furthermore, CAD does not allow for organizing one's files in a structured database.

Both CAD and SICAR do not allow raster editing; the operator cannot query pixels or optimize the image. Color data are crucial to characterize some types of artifacts, especially those with a decorated or painted surface, of which, for example, the specific conservation problems of each color should be analyzed.

CAD and SICAR do not allow for any interaction between raster and vector graphics, and each graphic survey is executed by manual drawing. The result is highly subjective, and each thematic mapping is different from any other, even if carried out by the same operator on the same photographic basis. Finally, restorers rarely use a unique system to compile their thematic maps. When dealing with canvases, painted tables, ceramics, fresco

paintings, and mosaics, restorers use, often empirically, diverse commercial software for vector graphics and image processing without adopting a standard methodology.

### 2.3. Geographic Information System

In more recent years, the development of low-cost and easy-to-use Geographic Information System software including spatial attributes and mapping elements, has made it possible to use this technology for non-geographic projects. From relatively large areas, GIS has been used on mobile artifacts; especially in Italy, Spain and Portugal, experimentation on the statistical analysis of the degradation of painted canvas and tables has started [12–15]. Two GIS systems were used for these experiments: QGIS®—free and open-source, and ArcGIS®, proprietary software of Esri®. In particular, QGIS® has proven to be widely used for scientific research in territorial, archaeological, and artistic history. For this software, significant plugins have been developed, among them the Semi-automatic Classification Plugin (Version 7.8.0) that combines multispectral images and raster analysis, allowing the operator-controlled semi-automatic classification of environmental remote sensing images and providing tools to accelerate the process of classification of soil areas [16].

In the archaeological field, a plugin called pyArchInit (Version 2.4.6.) has been developed for QGIS® that allows access to a global database server that can be consulted and modified-like PostgreSQL-favoring the homogeneity promoting the homogeneity of the solutions adopted and exporting projects through interactive systems that can be used on the web: the so-called web-GIS [17]. This plugin satisfies the archaeological community's growing need to computerize excavation documentation, but can also manage documentation in the architectural and historical-artistic fields.

### 2.4. Diagnostic Investigations

As far as diagnostic investigations are concerned, white balance and color correction, post-production, and interpretation modes are the main issues.

Digital diagnostic images require a colorimetric correction process, during both acquisition and post-production. In general, all the adjustment and balance operations greatly influence the interpretation of multispectral images as they modify the color tones expressed in the visible spectrum, corresponding to the wavelengths reflected by the material surface. In the case of induced luminescence images that give a response in the field of visible, such as ultraviolet-induced luminescence (UVL) and visible-induced visible luminescence (VIVL) images, artistic materials show different grades and tones of fluorescence according to their composition and aging. Therefore, white balance in shooting and color correction can substantially affect the fidelity and reproducibility of images, making interpretation inaccurate and comparisons between images taken at different times and with different settings inconsistent. As far as IR and UV reflected images and visible-induced infrared luminescence (VIL) images are concerned, white balance mostly impacts on the post-production of false-color reflected images, such as ultraviolet-reflected false color (UVRFC) and infrared-reflected false color (IRRFC).

To obtain consistent and comparable data, it is necessary to follow standards currently represented by the results of the Charisma project of the British Museum [18]. However, these standards are not easy to implement, and in practice, more readily available commercial colorimetric references are used, but they do not offer optimal results. Furthermore, commercially available cameras are designed to provide aesthetically pleasing images and not for scientific analysis of artifacts and, as a result, may make undesired changes to captured multispectral images. The automatic adjustments incorporated into the cameras include white balance adjustments, contrast, brightness, sharpness, Automatic Gain Control (AGC) and control of the dynamic range in low light situations. To solve these problems, in addition to the Charisma Project, several manuals and scientific articles have been published for the correct use of colorimetric references for multispectral imaging [19,20].

The interpretation of multispectral images changes depending on the technique used, the artifact under investigation, and its history dating. Concerning grayscale monochrome

images such as infrared-reflected (IRR) and ultraviolet-reflected (UVR), reading and interpretation difficulties depend exclusively on the state of preservation of the artifact under examination and the information searched, so contrast and opacity of the objects in the scene are the only elements on which the restorer and the diagnostician use to recognize the characteristics.

Concerning the phenomenon of reflectance of the materials presented in color diagnostic images, such as induced luminescence images and false-color reflected images, identifying the characteristics searched is more complicated. The characteristics are shown in areas of color that are more or less homogeneous and with variations in tonality and opacity that are always different. These variations depend on the artifact's execution techniques, the surface materials used and their aging. In addition, the materials, such as pigments and binders, are mixed and layered with each other, and the application techniques vary greatly depending on the historical period. As a result, materials react in multiple spectrum bands simultaneously with different levels of intensity. Therefore, the recognition of characteristics through reading the fluorescence phenomenon is subject to an interpretative process depending on the examiner's experience and the reference literature for any specific case.

Over the years, many scientific types of research have been conducted in support of the recognition of fluorescent materials. Among the main ones we can mention (a) a mathematical model based on the Kubelka-Munk theory that studies the pigment-binder interaction [21,22], (b) a false color imaging technique called ChromaDI that enhances in the visible image the differences between the optical behavior of the various pigments taking into account the changes that occur during the transition from short to longer wavelengths [23], and (c) a methodology to classify different pigments through Hyper Spectral Imaging (HSI) that acts in the Short Wavelength Infrared (SWIR) region [24].

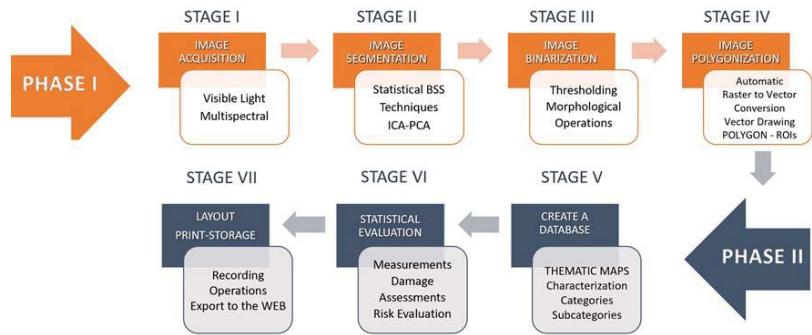
Regarding the statistical analysis of the multispectral image, the same techniques we use here have proved to be very efficient for improving the readability of ancient, degraded manuscripts and palimpsests [25]. Furthermore, in archaeology, these techniques have been used to reveal details otherwise invisible or difficult to discern on the surface of painted walls [26]. However, multispectral techniques are rarely combined with statistical image processing and have never been used to facilitate the restorer in creating thematic maps for documentation. Finally, it is central to specify that regardless of the type of technique and artifact investigated, archaeologists and conservators/restorers are the only ones who know the artifact's material characteristics and must always be involved in the reading interpretation of diagnostic investigations.

### 3. A New Methodological Approach

In the common documentation practice, the available diagnostic images are analyzed separately and their analysis is based on visual observation of the reflectance phenomena of materials, although in some cases they can be distorted and difficult to interpret. Our strategy instead considers the set of images available as a whole. The aim is to subdivide the painted area into regions through strategies of Blind Source Separation (BSS) [27], which have long been used in other areas, such as document image processing [28]. Then, we propose a rigorous and semi-automated analysis procedure for an easy, fast, and repeatable automatic polygonization of the extracted regions using standard software tools [29]. Therefore, the restorer's fundamental choice is the only remaining subjectivity, but it is based on a precise and repeatable set of objective and scientifically valid measurements. It is essential to point out that the restorer's role in directing the investigations and interpreting the images is central and cannot be delegated.

As illustrated in Figure 1, our methodology includes a first phase during which the diagnostic images are acquired and manipulated for analysis purposes. It includes four stages: image acquisition, image segmentation, threshold-based extraction of the regions of interest (ROI), and mapping from raster to vector representation. In the second phase, the methodology applies classification and analysis methods to determine the state

of conservation of the artwork. The resulting output is then archived according to the specified requirements.



**Figure 1.** Methodology lifecycle.

### 3.1. Stage 1: Acquisition

The methodology's efficacy is strongly affected by the quantity and quality of the digital images provided, representing the input for the following stages' processing and analysis algorithms. As a general rule, the main requirements of photographic reproduction for documentation are precision and consistency with respect to the original dimensions of the artwork, readability of all its parts, high accuracy in color reproduction, uniform illumination on the whole image, and absence of reflections that could impair the analysis. Regarding the respect of the accuracy of the digital reproduction of an artifact, the most used software tools in cultural heritage are Agisoft Fotoscan© and Archis-Siscan©; in particular, Archis is highly useful to the rectification of complex objects or architectures and allows for a two-dimensional graphic restitution of the artifact. Below is a list of the requirements necessary for the methodology we propose.

One of the problems presented by a quantitative approach to the study of artifacts through digital images lies in the strong dependence of obtainable results by changing scale of analysis: indeed, reducing spatial resolution, the information related to the morphology of material features present on the surface is lost. So, for a complete characterization of the survey, it becomes necessary to work on images with a correct spatial resolution, in proportion to the artifact's size and to the enlargement level on which the analysis will be carried out. Spatial resolution indicates the amount of detail visible in it. As for our methodology, the main advantage consists exactly in the great number of available details, as each detail represents further information. The methodology has been tested on images with different degrees of spatial resolution, we recommend using images with a spatial resolution of 300 pixels per inch—or a minimum of 150 pixels per inch—and 16 bits per channel.

Since the image processing techniques work on multiple images simultaneously, image data coherence is one of the fundamental requirements of the methodology. In order to proceed to the second phase concerning the application of segmentation algorithms, it is necessary that all the images acquired on the same artifact, in the same phase of the conservation/restoration intervention, are correctly registered with each other. Image registration is used to align images of the same subject taken with different acquisition techniques. Alignment involves eliminating slight rotations and tilts and re-sampling the images to the same scale.

It is not always possible to obtain full alignment of all acquired data because artifacts may change their morphology during the restoration intervention. In these cases, it is essential to provide consistency between sets of data acquired during the same phase of the intervention. Reference [30], dating 2003, aims to present a review of recent and classical

image registration methods. More recently, a new image registration framework has been proposed in [31], based on multivariate mixture model (MvMM) and neural network estimation. Reproducibility of the shooting session must be guaranteed by filling out an activity report and a technical diagnostic report, to be kept both to verify the methodology correctness and for further comparisons in time. Reproducibility should also be ensured by saving, cataloguing and archiving all the original files and the connected meta-data, relating to all work phases.

The choice of the spectral range for image acquisition affects the analysis process, since each range is associated with different quantitative and qualitative information. Concerning the choice of the diagnostic acquisition technique to be applied, this must be assessed by conservators/restorers or by researchers, and will have to be aimed at providing answers to specific (previously expressed) problems, and to questions concerning artifacts conservation and restoration, in their cultural and technological framework. In the absence of multispectral images, the methodology can be applied to images only taken in the visible spectrum. Reference [32] describes an experimental approach to use “color” uniformity alone as a criterion for dividing the image into disjoint regions of interest corresponding to distinguishable features on the surface of the artifact. In detail, this approach was used to highlight overwritten text in a palimpsest, showing that satisfactory results can be obtained with a method of color decorrelation even starting from visible-light images. The results can often be as highly discriminative as those provided by diagnostic and multispectral images. In this case, the color space’s choice is central, as different spaces represent color information in different ways; see, e.g., RGB, HSV, and L\*A\*B\* channels [33,34]. In our tests, the HSV color model has proven to be very useful for color segmentation in complex contexts such as artistic artifacts’ images. The main reason for this usefulness is in V’s components, i.e., in the values corresponding to brightness gradations, which allow to detect even the slightest variations in light intensity, and therefore the smallest discontinuities between areas of interest. Color properties thus described have an immediate perceptive interpretation by conservators/restorers, who are accustomed to distinguish colors according to their visual perception, on which this model is based.

### 3.2. Stage 2: Segmentation

Segmentation is the process of grouping spatial data into multiple homogeneous areas with similar properties. In our case, the properties (or features) we consider are the spectral responses of the different materials, that is, the local reflectance values measured in all the available channels. The regions of interest (ROI) we want to distinguish, segment, and extract are the pixel sets with homogeneous features (such as locations, sizes, and color), which correspond to parts of the artifact made of similar materials. Unfortunately, regions showing different features typically overlap with one another in all the channels, because typically, the materials are mixed and stratified with each other. This often makes segmentation and ROI extraction difficult. The visual inspection performed by diagnosticians or conservators-restorers can be very complicated, time-consuming, or even impossible, particularly when just slightly different spectral characteristics must be distinguished from many channels. Thus, this task can be performed more efficiently and objectively through automated image analysis techniques. In particular, manipulating the input channels to produce a number of maps, each showing a single or a few ROIs, can significantly facilitate the segmentation. Mathematically, this would be accomplished easily if the different materials’ spectral emissions were known, but this is seldom the case. To extract the different regions from multispectral data with no knowledge of their spectra, some assumptions must be made on the regions themselves and the mixing mechanism that produces their spectral appearance. For the mechanism, we assume an instantaneous linear model with  $M$  hyperspectral channels and  $N$  distinct features

$$x_i(t) = \sum_{j=1}^N a_{ij} s_j(t), \quad i = 1, \dots, M \quad (1)$$

where  $x_i(t)$  is the value of the data at channel  $i$  and at pixel  $t$ ,  $a_{ij}$  is the spectral emission of the  $j$ -th feature in the  $i$ -th channel, and  $s_j(t)$  is the value of the  $j$ -th feature at pixel  $t$ . Notice that an additional assumption in this model is that the spectra  $a_{ij}$  are assumed to be uniform all over the image. This model is also called instantaneous because the data values at each pixel only depend on the feature values in that pixel and not on any neighborhood of it. If we are able to extract the map  $s_j(t)$  from the data  $x_i(t)$ , then it will be easy to extract the ROIs related to the  $j$ -th feature by just locating the regions where  $s_j(t)$  assumes significant values. Extracting  $s_j(t)$  from  $x_i(t)$  with no knowledge of  $a_{ij}$ , is a blind source separation problem (BSS), which can only be solved by further assumptions on  $s_j$ . In particular, statistical BSS techniques such as principal component analysis (PCA) [27], and independent component analysis (ICA) [35] reasonably assume that the different features  $s_j$  have some degree of statistical independence. Indeed, as the patterns formed by different materials in the painted surface are likely to be independent of one another, it is also likely that their central mutual statistics nearly vanish all over the images, that is, assuming zero-mean feature maps:

$$\langle s_k^\alpha \cdot s_l^\beta \rangle \simeq 0 \quad \forall k \neq l \quad (2)$$

where  $\alpha$  and  $\beta$  are arbitrary integers, and the angle brackets denote statistical expectation.

Particular cases of (2) are zero-correlation ( $\alpha = \beta = 1$ ), leading to PCA and other second-order approaches, and statistical independence, i.e., (2) is true for all  $\alpha$  and  $\beta$ , leading to ICA. By these assumptions, the result is obtained by minimizing the following summation with respect to all the  $s_j$ :

$$\sum_{k \neq l} |\langle s_k^\alpha \cdot s_l^\beta \rangle| \quad \text{subject to } x_i(t) = \sum_{j=1}^N a_{ij} s_j(t) \quad (3)$$

If the preliminary assumptions are satisfied, this produces a new set of images, each depicting one and only one of the desired feature maps, that is, something approximately proportional to  $s_j$ . By estimating matrix  $\{a_{ij}\}$ , both PCA and ICA estimate the features  $s_j$  by combining linearly the normally correlated  $x_i$  to produce a different set of images that are uncorrelated or statistically independent.

Since each output map assumes significant values only where a single feature is present, the ROIs characterized by such a feature can be extracted by just distinguishing between foreground and background. In fact, at best, two primary gray levels dominate each output channel, and only a specific ROI is highlighted. All pixels in the ROI will have similar gray values, and the rest of the image will get confused in the background.

### 3.3. Stage 3: Binarization

Since the purpose of the methodology is to obtain a precise vector polygon for each classified ROI, the third stage consists of eliminating the overabundance of information caused by the average gray levels, which cause the typical ramp edges between the area of interest and the background. Gradients, in this case, can be classified as noise that slows down and complicates the process of identifying ROIs. To further segment the ROIs from the background, we use a simple recursive thresholding algorithm [36].

Given the output channels of the second stage, which correspond to images  $f(x, y)$  in grayscale, a gray gradation is fixed, called intensity threshold  $T$ . In the binary output image, the pixels labeled with 1 are called object points, while those set to 0 are the background points. The segmentation outcome is strongly influenced by the choice of the threshold  $T$ , which can either be constant throughout the image (global thresholding) or vary dynamically from pixel to pixel (local thresholding).

$$g(x, y) = \begin{cases} 1, & \text{if } f(x, y) \geq T \\ 0, & \text{if } f(x, y) < T \end{cases} \quad (4)$$

Global thresholding fails in the cases where the image content is not evenly illuminated. Hence, it is essential to respect the lighting consistency during the shooting stage to avoid further optimization pre-processing. The  $T$  value can be chosen canonically, using the average value in the grayscale, but in many cases the restorer's choice is made by trial and error, that is, testing different values to determine one that makes the output satisfactory [37]. This can be done on a single channel and adapted to the others, or by choosing a  $T$  value for each channel. In any case, each selected value must be included in the report accompanying the thematic map to guarantee the results' reproducibility.

It is worth observing that the choice of the output image to work with is up to the restorer. This choice is needed to filter only the ROIs for the specific thematic map to be generated; in fact, a fully automatic execution would lead to the identification and selection of undesired regions. The final result is a set of binary masks that identify the ROIs by step edges that will facilitate the subsequent stages. In rare cases, the extracted masks may still have noise pixels inside or outside the ROIs. These pixels need to be removed in order not to interfere with the subsequent polygonization processes. For this purpose, basic morphological operations such as region filling, thinning, and thickening can be used to clean noise and modify regions [38].

Alternative approaches to the generation of the binary masks use neural networks, such as, e.g., the Kohonen self-organizing map (SOM) [39,40], which is based on competitive training algorithms [41]. The advantage of SOM is preserving the input samples' topology [42]. We tested this type of neural network on different types of artifacts, but they proved to be very useful only in some cases, when the ROIs are already visible in all diagnostic images, that is, in cases of easy segmentation. Moreover, the extraction of binary masks with this method proved to be too time-consuming and not easy to use for non-experts. For this reason, we decided not to include them in the final methodology proposed here.

### 3.4. Stage 4: Polygonization

Raster to vector data conversion is a central function in GIS image processing and remote sensing (RS) for data integration between RS and GIS [43]. In general, there are two types of algorithms, namely, vectorization of lines and vectorization of polygons; only the latter is used in our methodology. This function creates vector polygons for all connected regions sharing a communal pixel value [29]. The precision of the mapping depends on several factors, including the spatial complexity of the images. As the resolution of an image becomes sharper, the data volume increases; for this reason, it is important to perform the pre-processing operation of stage 3 to classify information, clean the shapes and the edges of the ROIs, and ensure their topological coherence.

### 3.5. Methodology in QGIS

In the first phase of the methodology, the raster images are inserted into QGIS using a metric reference system, such as WGS84 EPSG:4326, and associating a worldfile to each image. A worldfile is a collateral file of six plain text lines used by geographic information systems to georeference raster map images. The file specification was introduced by Esri® and consists of six coefficients of a similar transformation that describes the position, scale, and rotation of a raster on a map. This procedure allows us to have the starting data consistent with each other, geometrically correct and divided into different layers according to the acquisition mode.

The second stage involves the analysis of all diagnostic images simultaneously. The PCA algorithm implemented in QGIS is adequate for our purposes—see the Processing Tools for Raster Analysis, in GRASS tools (*i.pca*), while ICA would require new modules to be implemented in Python. Alternatively, both algorithms can be implemented entirely in Matlab (Version 9.7 R2019b, MathWorks, Natick, MA, USA) [44].

The third stage of binarization is performed individually on each output image. The conservator-restorer chooses among the outputs the ones that best fulfil their cognitive

needs. The thresholding algorithm and the morphological functions are present in QGIS, in the list of processing tools of raster analysis, classified for layers or tables. Then, QGIS can perform raster to vector conversion of each binary mask extracted in the third stage. The tool is found in the main menu, in section Raster, Conversion, Polygonization (from Raster to Vector). This conversion occurs very quickly even with complex files and allows for the creation of a vector polygon that faithfully reflects the edges of areas of interest recognized and highlighted by diagnostic investigations.

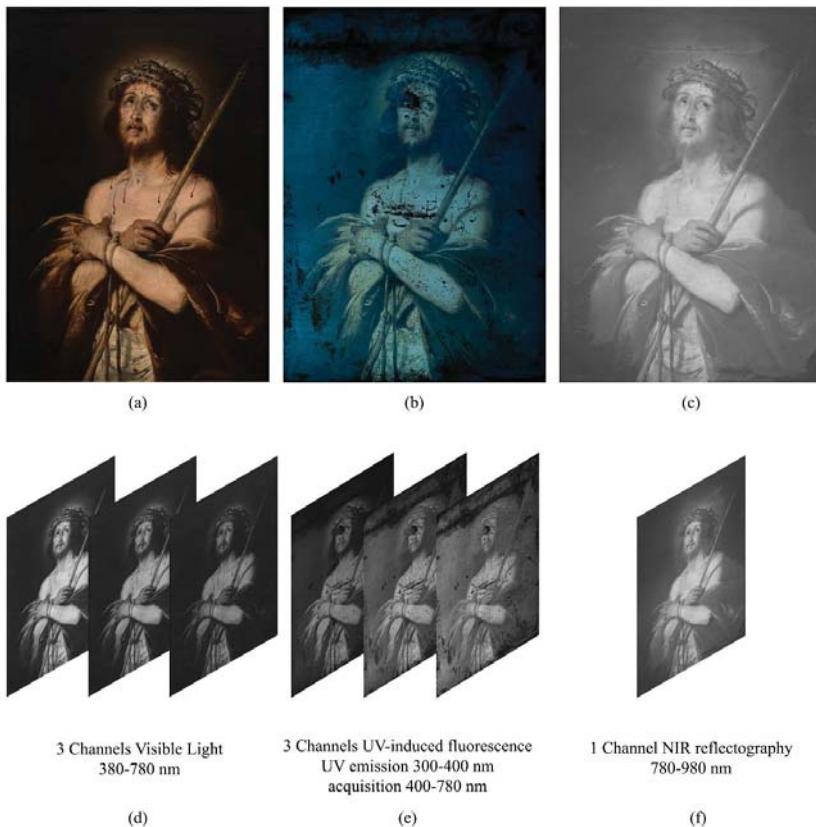
All the stages of the second phase, related to the operation of classification, analysis, and storage, are performed using computing tools of QGIS to create a database supporting measurements and statistical analysis. Hence, the output is a detailed and personalized description of each polygonal space created, possibly in different file formats, according to the documentation's needs. Each extracted polygon is classified into categories and subcategories through an attribute table, which favors different analysis types, such as damage assessments and risk evaluation, which can be conducted quantitatively and objectively.

#### 4. Case Study

The methodology's effectiveness is demonstrated in the specific case of a canvas painting by showing that the thematic map that is typically extracted manually can be derived through the stages described above. The chosen artwork is *Ecce Homo* by Bernardo Strozzi, an oil painting on canvas, 1620–1622, in size  $105 \times 75 \text{ cm}^2$ , which is in good conservation status. It underwent a restoration in recent times, including a cleaning of the superficial paint layer and the removal of the old restoration interventions; subsequently, a pictorial retouch with paint was carried out. For all phases of our methodology, apart from stage 2 that requires MatLab's use for the segmentation algorithms, we used the open-source software QGIS (version 3.10.2-A Coruña, whit Grass 7.8.2.).

By following stage 1 of the proposed method, the painting was captured in three different modalities, under visible light illumination (Figure 2a), UV-fluorescence (Figure 2b) and Near-Infrared Reflectography 780–980 nm (Figure 2c). The fluorescence image was subtracted of the visible stray light to highlight the regions that really produce fluorescence under ultraviolet illumination (see [19,45] for details). In the second stage, the three images were processed by PCA and ICA; the output images are shown in Figure 3a,b. A further attempt with ICA has been made on a subset of six channels, obtained excluding the infrared and resulting in the outputs of Figure 3c. It is essential to make explicit that every image produced by the statistical processing no longer corresponds to a specific wavelength range of, but is a recombination of their intensities, highlighting one or more of the ROIs required, which appear in different gray levels. The 20 output images in Figure 3a–c are the new data set that the restorer can inspect for study and feature recognition.

After inspecting the output images, the restorer chose those where the significant regions are most noticeable. These images, reported in Figure 4a,e,i, show some features related to the materials used: Figure 4a highlights the pictorial retouches performed on the background and Figure 4e highlights the pictorial retouches performed on the body of Christ. These two features in the original data were visually superimposed in the same spectral channel, while in this phase they are separated in different outputs, despite having been operated at the same time. Therefore, we can say that segmentation is due to the different pigments used for retouching. In fact, the figure of Christ has likely been restored using titanium white pigment, and the background has been restored using varnish colors. Even without (destructive) chemical analysis or any other more specialized technique, image processing has allowed us to distinguish between regions that look similar in the acquired channels. Analogously, in Figure 4i-l, the red lacquer used for the blood of Christ is highlighted.



**Figure 2.** Phase I, stage 1—Ecce Homo by Bernardo Strozzi, oil on canvas, 1620–1622,  $105 \times 75 \text{ cm}^2$ : (a) Standard RGB; (b) UV-induced fluorescence; (c) NIR reflectography; (d–f) Respective spectral Channels of the acquired images. Images captured by Paolo Triolo, under permission of the Ministry of Cultural Heritage and Activities and Tourism, National Gallery of Palazzo Spinola, Genova, Italy.

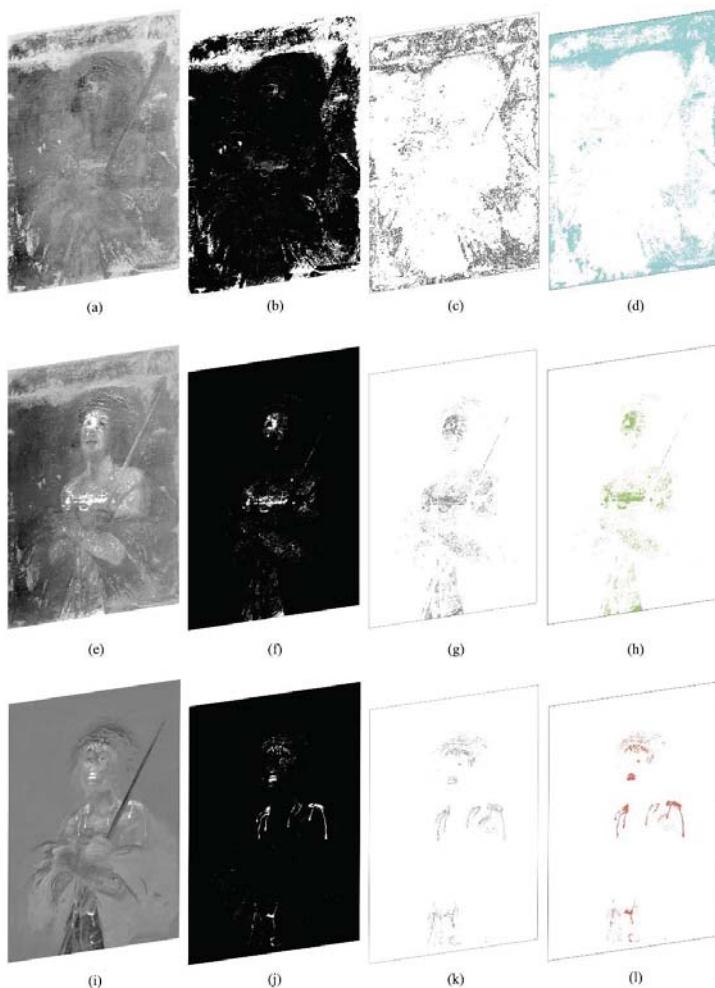
The third stage of the methodology, the creation of binary masks through the threshold algorithm, was only carried out on the three outputs chosen by the restorer. In our case, in the range 0–255, we chose a threshold  $T = 180$ . Our result is shown in Figure 4b,f,j. For stage 3, we have used the QGIS Plugin Value tool to choose the threshold value, and the raster analysis processing tool Classified for layers or tables for creating the binary masks.

The fourth stage consists of the graphic design's automatic extraction to create the thematic map. The binary masks shown in Figure 4c,g,k have been transformed from raster to vector drawings, thus obtaining automatically closed vector polygons that comply with topographical rules of adjacency and overlap. For stage 4, we have used the QGIS polygonizer tool from raster to vector.

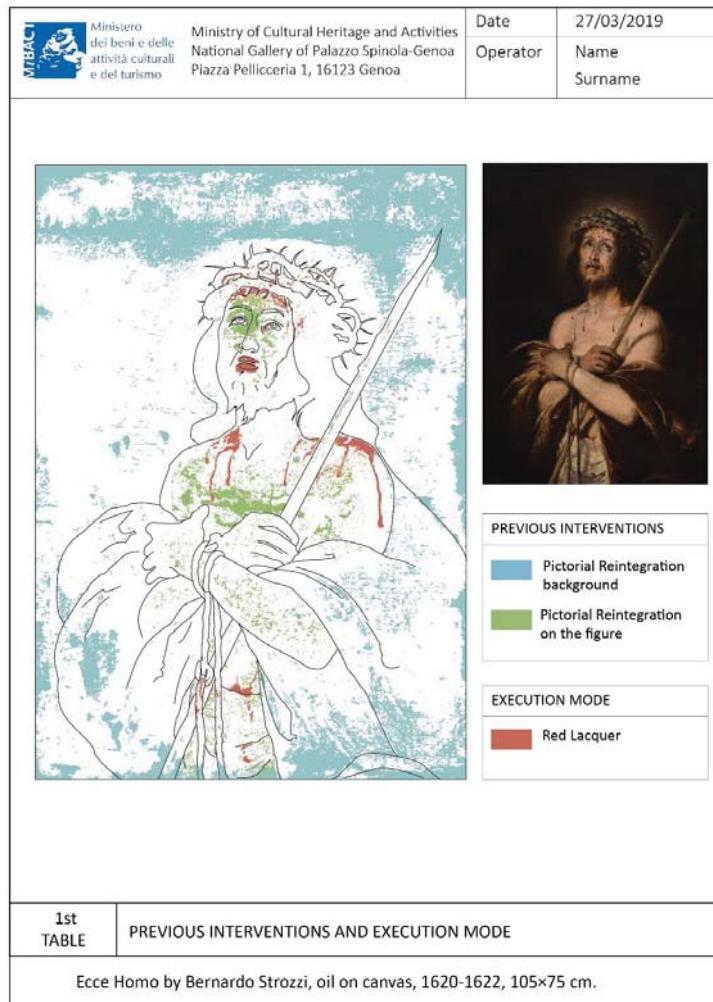


**Figure 3.** Phase I, stage 2: (a) Output channels obtained by principal component analysis (PCA) from the entire data set in Figure 2; (b) Output channels obtained by independent component analysis (ICA) from the entire data set in Figure 2; (c) Output channels obtained by ICA from the multispectral cube with no IR data in Figure 2a,b,d,e.

In the second-phase stages, each extracted polygon is then classified in a corresponding ROI layer and characterized by a different color and texture, see, to create the legend in the thematic map, see Figure 4d,h,l and Figure 5. In particular, for stage 5, we used the characterization in the QGIS Layer Style tool. The extracted polygons corresponding to the ROI have been estimated as a percentage of the total area of interest, dividing them by Layers and associating them into a Table of Attributes. The topological relationship between the database and the graphics is that it is possible to query the data directly by querying the graphic design and automatically exporting the legend and the statistical analysis results in the printing layout phase. QGIS has a useful and versatile Layout Window supporting creating complex sheets and can save templates to be reused in the future. The metadata have been included in the layout, which contains the institution's logo, the author's name, the dimensional characteristics of the object, the table number, the date, the documentation operator, and a legend (or glossary) linked to the drawing. This type of metric result is useful for monitoring, conservation, and restoration, see Figure 5. For this case study, we chose to show only the front of the artifact as the back and side profile of the canvas did not have any interesting feature.



**Figure 4.** Phase I stages 3–4: (a,e,i) Images processed from the previous stage (in Figure 3) and chosen to identify regions of interest (ROIs); (b,f,j) Corresponding binarized versions; (c,g,k) Image polygonization, raster to vector conversion. Phase II stage 5: (d,h,l) Characterization of the extracted polygons.



**Figure 5.** Thematic map-specific colors are assigned to the ROIs (statistical data are omitted).

## 5. Conclusions

We examined some of the problems concerning the graphic documentation in cultural heritage, i.e., the difficulty in analyzing the diagnostic images, the excessive subjectivity and approximation of the transcription of the relevant information, the complexity, and the long time needed to transcribe information through manual procedures. These problems lead the restorers to choose only the essential information to document, thus making the graphic documentation incomplete and far from guaranteeing reproducibility. There is currently no official or de facto methodological standard that considers all the possibilities offered by image processing and scientific visualization, and commercially available software tools. Consequently, we propose a semi-automated methodology to facilitate and improve the diagnostic investigation while reducing drastically manual interventions. The result of its application is an objective, formal, and accurate graphic documentation to plan restoration, monitoring, and conservation interventions.

Moreover, as a novelty in this field, image segmentation algorithms have demonstrated their potential to reduce subjectivity and accelerate the entire process. The time needed for the entire methodology to be applied can be evaluated according to two factors. The first is the characteristic and speed of the device in use. Basically, all the algorithms used require a short calculation time ranging from a minimum 4–5 s and a maximum of 1–5 min. These times were evaluated considering a spatial resolution of 300 dpi and a low/medium power device. The power of the device’s graphics card and the massive number of images to be examined are the only two factors that can increase the computation time of the BSS algorithms. The second factor includes both the operator’s ability to use the software and the number of thematic maps to be performed. The application of the entire methodology requires a medium/advanced knowledge of the software mentioned. Furthermore, the timing of the creation of the thematic maps depends on the reasoning time of the user themselves and on the features to extract and document.

Based on image analysis processes, the methodology can be applied to any surface, regardless of the spatial complexity of the object or its extent. However, in order to obtain real data, it is necessary that the photographic reproduction of the artefact respects the real dimensions, with the minimum margin of error.

To date, QGIS has been assessed as the most appropriate tool to support each step of the methodology, as in this framework the operator has all the necessary image analysis tools, combining high potential and ease of use. In a previous, empirical and preliminary study preceding the formal methodology [46] we tested the BBS algorithms coupled to neural networks. As future work, we plan to investigate additional algorithms, and in order to ensure a more general applicability of the results we expect to replicate the experiment on a larger number of case studies and the implementation of well-known blind methods to assess the reliability and stability of the results achieved. Another future development could be to integrate our method with some recent experiences and advanced strands of research that are trying to overcome some of the limitations of documentation, offering web-based solutions/platforms able to perform the operations of survey (mapping) of conservation, restoration and preservation in a single environment/system; also by exploiting three-dimensional models [47–50].

**Author Contributions:** Conceptualization, A.A. (Annamaria Amura), A.A. (Alessandro Aldini) and A.T.; methodology, A.A. (Annamaria Amura), S.P., E.S., A.T. and P.T.; software, A.A. (Annamaria Amura), and A.T.; validation, A.A. (Annamaria Amura), A.T. and P.T.; formal analysis, E.S. and A.T.; data curation, P.T. and A.A. (Annamaria Amura); writing—original draft preparation, A.A. (Annamaria Amura); writing—review and editing, A.A. (Annamaria Amura), A.A. (Alessandro Aldini) and E.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data used to produce the results shown here have been acquired by one of the authors (P.T.) under permission of the Ministry of Cultural Heritage and Activities and Tourism, National Gallery of Palazzo Spinola, Genova, Italy, which maintains all the property rights.

**Acknowledgments:** The authors are grateful to the QGIS Italia community for their technical support in adapting the software to the needs of the methodology presented.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. *Normal 17/84: Elementi Metrologici e Caratteristiche Dimensionali: Determinazione Grafica., in Raccomandazioni Normal. Alterazioni dei Materiali Lapidei e Trattamenti Conservativi: Proposte per L'unificazione dei Metodi Sperimentali di Studio e di Controllo;* Consiglio Nazionale delle Ricerche (CNR), Istituto Centrale per il Restauro (ICR): Roma, Italy, 1984.
2. Sacco, F. *Sistematica Della Documentazione e Progetto di Restauro;* Boll. ICR. Nuova Ser. N.4; ICR: Roma, Italy, 2002; pp. 28–50.

3. Bezzi, L.; Bezzi, A. Proposta per un metodo informatizzato di disegno archeologico Il disegno archeologico. In Proceedings of the ArcheoFOSS. Open Source, Free Software e Open Format nei Processi di Ricerca Archeologici, Foggia, Italy, 5–6 May 2010; De Felice, G., Sibilano, M.G., Eds.; Edipuglia: Puglia, Italy, 2011; pp. 113–123.
4. Agosto, E.; Ardissono, P.; Bornaz, L.; Dago, F. 3D Documentation of Cultural Heritage: Design and Exploitation of 3D Metric Surveys. In *Applying Innovative Technologies in Heritage Science*; George Pavlidis (Athena—Research and Innovation Center in Information, Communication and Knowledge Technologies, Greece); IGI Global: Hershey, PA, USA, 2020; pp. 1–15.
5. ICCD-Istituto Centrale per il Catalogo e la Documentazione. *La Documentazione Fotografica Delle Schede di Catalogo: Metodologie e Tecniche di Ripresa*; Galasso, R., Giffi, E., Eds.; ICCD: Roma, Italy, 1998; ISBN IEI0127676.
6. International Standard Organization (ISO). ISO/3567-1:1 1998. *Technical Product Documentation*; Organization and Naming of Layers for CAD. Part 1: Overview and Principles; International Standard Organization (ISO): Geneva, Switzerland, 1998.
7. Eiteljorg, H., II. Documentation with CAD. In *Boll. Dell’Istituto Cent. per Restauro*; ICR: Roma, Italy, 2002; Volume 5, pp. 45–50.
8. ISO 13567-1 Technical Product Documentation. Organization and Naming of Layers for CAD. Part 1: Overview and Principles. 1998. Available online: <https://www.iso.org/obp/ui/#iso:std:iso:13567-1:ed-1:en> (accessed on 10 March 2021).
9. Rinaudo, F.; Eros, A.; Ardissono, P. Gis and Web-Gis, Commercial and Open Source Platforms: General Rules for Cultural Heritage Documentation. In Proceedings of the XXI International CIPA Symposium, Athens, Greece, 1–6 October 2007.
10. Rajcic, V. Risks and resilience of cultural heritage assets. In *Proceedings of the SBE 16 Malta Europe and the Mediterranean Towards a Sustainable Built Environment*; Borg, R.P., Gauci, P., Staines, C.S., Eds.; SBE Malta: Msida, Malta, 2016; pp. 325–334.
11. Fabiani, F.; Grilli, R.; Musetti, V. Verso nuove modalità di gestione e presentazione della documentazione di restauro: SICaR web la piattaforma in rete del Ministero dei Beni e delle Attività Culturali e del Turismo. *Boll. Ing. Coll. Degli Ing. Della Toscana* **2016**, *3*, 3–13. [[CrossRef](#)]
12. Baratin, L.; Bertozzi, S.; Moretti, E.; Saccuman, R. GIS applications for a new approach to the analysis of panel paintings. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; LNCS; Springer: Berlin/Heidelberg, Germany, 2016; Volume 10058, pp. 711–723. [[CrossRef](#)]
13. Fuentes-Porto, A. La Tecnología Sig Al Servicio De La Cuantificación Numérica Del Deterioro En Superficies Pictóricas. Un Paso Más Hacia La Objetivización De Los Diagnósticos Patológicos. In Proceedings of the V Congreso Grupo Español del IIC. Patrimonio Cultural, Criterios de Calidad en Intervenciones, Madrid, Spain, 18–20 April 2012; GE publicaciones: Madrid, Spain, 2012; pp. 363–369.
14. Henriques, F.; Gonçalves, A.; Bailão, A. Tear feature extraction with spatial analysis: A thangka case study. *Estud. Conserv. Restauro* **2013**, *1*, 10–23. [[CrossRef](#)]
15. Henriques, F.; Gonçalves, A. Analysis of lacunae and retouching areas in panel paintings using landscape metrics. In *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*; LNCS; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6436, pp. 99–109. [[CrossRef](#)]
16. Congedo, L.; Macchi, S. *Semi-Automatic Plugin Classificazione per QGIS*; ACC Dar: Roma, Italy, 2013; Available online: <http://www.planning4adaptation.eu/> (accessed on 10 March 2021).
17. Gugnali, S.; Mandolesi, L.; Drudi, V.; Miulli, A.; Maioli, M.G.; Frelat, M.A.; Gruppioni, G. Design and implementation of an open source G.I.S. platform for management of anthropological data. *J. Biol. Res.* **2012**, *85*, 350–353. [[CrossRef](#)]
18. Dyer, J.; Verri, G.; Cupitt, J. (Eds.) *Multispectral Imaging in Reflectance and Photo-Induced Luminescence Modes: A User Manual*, 1st ed.; Online: European CHARISMA Project.: Web publication/Site; British Museum: London, UK, 2013.
19. Triolo, P.A.M. *Manuale Pratico di Documentazione e Diagnostica per Immagine per i BB.CC*; Il Prato: Padova, Italy, 2019.
20. Verri, G.; Clementi, C.; Comelli, D.; Cather, S.; Piqué, F. Correction of ultraviolet-induced fluorescence spectra for the examination of polychromy. *Appl. Spectrosc.* **2008**, *62*, 1295–1302. [[CrossRef](#)]
21. Verri, G.; Comelli, D.; Cather, S.; Saunders, D.; Piqué, F. Post-capture data analysis as an aid to the interpretation of ultraviolet-induced fluorescence images. In *SPIE 6810 Computer Image Analysis in the Study of Art*; Stork, D.G., Coddington, J., Eds.; SPIE: Bellingham, WA, USA, 2008; Volume 6810, pp. 1–12. [[CrossRef](#)]
22. Zhao, Y. *Image Segmentation and Pigment Mapping of Cultural Heritage Based on Spectral Imaging*; Rochester Institute of Technology: New York, NY, USA, 2008.
23. Legnaioli, S.; Lorenzetti, G.; Cavalcanti, G.H.; Grifoni, E.; Marras, L.; Tonazzini, A.; Salerno, E.; Palleschi, P.; Giachi, G.; Palleschi, V. Recovery of archaeological wall paintings using novel multispectral imaging approaches. *Herit. Sci.* **2013**, *1*, 33. [[CrossRef](#)]
24. Capobianco, G.; Prestileo, F.; Serranti, S.; Bonifazi, G. Application of hyperspectral imaging for the study of pigments in paintings. In Proceedings of the 6th International Congress “Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin, Athens, Greece, 22–25 October 2013; Available online: [https://www.researchgate.net/profile/Giuseppe-Capobianco-2/publication/261223438\\_APPLICATION\\_OF\\_HYPERSPECTRAL\\_IMAGING\\_FOR\\_THE\\_STUDY\\_OF\\_PIGMENTS\\_IN\\_PAINTINGS/links/0a85e533a74e02f2fb000000/APPLICATION-OF-HYPERSPECTRAL-IMAGING-FOR-THE-STUDY-OF-PIGMENTS-IN-PAINTINGS.pdf](https://www.researchgate.net/profile/Giuseppe-Capobianco-2/publication/261223438_APPLICATION_OF_HYPERSPECTRAL_IMAGING_FOR_THE_STUDY_OF_PIGMENTS_IN_PAINTINGS/links/0a85e533a74e02f2fb000000/APPLICATION-OF-HYPERSPECTRAL-IMAGING-FOR-THE-STUDY-OF-PIGMENTS-IN-PAINTINGS.pdf) (accessed on 10 March 2021).
25. Salerno, E.; Tonazzini, A.; Bedini, L. Digital image analysis to enhance underwritten text in the Archimedes palimpsest. *Int. J. Doc. Anal. Recognit.* **2007**, *9*, 79–87. [[CrossRef](#)]
26. Salerno, E.; Tonazzini, A.; Grifoni, E.; Lorenzetti, G.; Legnaioli, S.; Lezzerini, M.; Marras, L.; Pagnotta, S.; Palleschi, V. Analysis of Multispectral Images in Cultural Heritage and Archaeology. *J. Appl. Spectrosc.* **2014**, *1*, 22–27. Available online: <http://www.alslab.net/wp-content/uploads/2018/03/JALS-1-art.4.pdf> (accessed on 10 March 2021).

27. Cichocki, A.; Amari, S. *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*; Wiley: New York, NY, USA, 2002; ISBN 978-0-471-60791-5.
28. Tonazzini, A.; Bedini, L.; Salerno, E. Independent component analysis for document restoration. *J. Doc. Anal. Recognit.* **2004**, *7*, 17–27. [[CrossRef](#)]
29. Junhua, T.; Fahui, W.; Yu, L. An Efficient Algorithm for Raster-to-Vector Data Conversion. *Geogr. Inf. Sci.* **2008**, *14*, 54–62.
30. Zitová, B.; Flusser, J. Image registration methods: A survey. *Image Vis. Comput.* **2003**, *21*, 977–1000. [[CrossRef](#)]
31. Luo, X.; Zhuang, X. MvMM-RegNet: A New Image Registration Framework Based on Multivariate Mixture Model and Neural Network Estimation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; LNCS; Springer Science and Business Media Deutschland GmbH: Berlin, Germany, 2020; Volume 12263, pp. 149–159. ISBN 9783030597153.
32. Salerno, E.; Tonazzini, A. Extracting erased text from palimpsests by using visible light. In Proceedings of the 4-th Int. Congress Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin, Cairo, Egypt, 6–8 December 2009; Ferrari, A., Ed.; Associazione Investire in Cultura–Fondazione Roma Mediterraneo: Roma, Italy, 2010; Volume II, pp. 532–535.
33. Jurio, A.; Pagola, M.; Galar, M.; Lopez-Molina, C.; Paternain, D. A comparison study of different color spaces in clustering based image segmentation. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*; Springer: Berlin/Heidelberg, Germany, 2010; Volume 81, pp. 532–541. [[CrossRef](#)]
34. Stigell, P.; Miyata, K.; Hauta-Kasari, M. Wiener estimation method in estimating of spectral reflectance from RGB images. *Pattern Recognit. Image Anal.* **2007**, *17*, 233–242. [[CrossRef](#)]
35. Hyvärinen, A.; Karhunen, J.; Oja, E. *Independent Component Analysis*; John Wiley & Sons, Inc.: New York, NY, USA, 2001; ISBN 047140540X.
36. Wu, B.; Chen, Y.; Chiu, C. Recursive Algorithms for Image Segmentation Based on a Discriminant Criterion. *World Acad. Sci. Eng. Technol. Int. J. Comput. Inf. Eng.* **2007**, *1*, 833–838.
37. Singh, S.; Talwar, R. Performance analysis of different threshold determination techniques for change vector analysis. *J. Geol. Soc. India* **2015**, *86*, 52–58. [[CrossRef](#)]
38. Shih, F.Y. *Image Processing and Mathematical Morphology. Fundamentals and Applications.*; CRC Press, Inc.: Boca Raton, FL, USA, 2009; ISBN 978-1-4200-8943.
39. Kohonen, T. The Self-Organizing Map. *Proc. IEEE* **1990**, *78*, 1464–1480. [[CrossRef](#)]
40. Koh, J.; Suk, M.; Bhandarkar, S.M. A multilayer self-organizing feature map for range image segmentation. *Neural Netw.* **1995**, *8*, 67–86. [[CrossRef](#)]
41. Yeo, N.C.; Lee, K.H.; Venkatesh, Y.V.; Ong, S.H. Colour image segmentation using the self-organizing map and adaptive resonance theory. *Image Vis. Comput.* **2005**, *23*, 1060–1079. [[CrossRef](#)]
42. Azorin-Lopez, J.; Saval-Calvo, M.; Fuster-Guillo, A.; Mora-Mora, H.; Villena-Martinez, V. Topology preserving self-organizing map of features in image space for trajectory classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2015; Volume 9108, pp. 271–280. [[CrossRef](#)]
43. Zhou, G.; Pan, Q.; Yue, T.; Wang, Q.; Sha, H.; Huang, S.; Liu, X. VECTOR AND RASTER DATA STORAGE BASED ON MORTON CODE. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *XLII-3*, 2523–2526. [[CrossRef](#)]
44. Hyvärinen, A. Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. *IEEE Trans. Neural Netw.* **1999**, *10*, 626–634. [[CrossRef](#)] [[PubMed](#)]
45. Triolo, P.A.M. Quattro opere a confronto. Interpretazioni e confronti delle ana-lisi di diagnostica per immagine sulle tele di Bernardo Strozzi. In *Bernardo Strozzi. Allegoria della Pittura*; Zanelli, G., Ed.; Sagep: Genova, Italy, 2018.
46. Amura, A.; Tonazzini, A.; Salerno, E.; Pagnotta, S.; Palleschi, V. Color segmentation and neural networks for automatic graphic relief of the state of conservation of artworks. *Color Cult. Sci. J.* **2020**, *12*, 7–15.
47. Grilli, E.; Remondino, F. Classification of 3D digital heritage. *Remote Sens.* **2019**, *11*, 847. [[CrossRef](#)]
48. Apollonio, F.I.; Basilissi, V.; Callieri, M.; Dellepiane, M.; Gaiani, M.; Ponchio, F.; Rizzo, F.; Rubino, A.R.; Scopigno, R. A 3D-centered Information System for the documentation of a complex restoration intervention. *J. Cult. Herit.* **2018**, *29*, 89–99. [[CrossRef](#)]
49. Apollonio, F.I.; Gaiani, M.; Bertacchi, S. Managing cultural heritage with integrated services platform. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*; Copernicus GmbH: Göttingen, Germany, 2019; Volume 42, pp. 91–98.
50. Soler, F.; Melero, F.J.; Luzón, M.V. A complete 3D information system for cultural heritage documentation. *J. Cult. Herit.* **2017**, *23*, 49–57. [[CrossRef](#)]





Article

# Computer Vision Meets Image Processing and UAS PhotoGrammetric Data Integration: From HBIM to the eXtended Reality Project of Arco della Pace in Milan and Its Decorative Complexity

Fabrizio Banfi \* and Alessandro Mandelli

Architecture, Built Environment and Construction Engineering (ABC) Department, Politecnico di Milano, 20133 Milano, Italy; alessandro.mandelli@polimi.it

\* Correspondence: fabrizio.banfi@polimi.it

**Abstract:** This study aims to enrich the knowledge of the monument Arco della Pace in Milan, surveying and modelling the sculpture that crowns the upper part of the building. The statues and the decorative apparatus are recorded with the photogrammetric technique using both a terrestrial camera and an Unmanned Aerial Vehicle (UAV). Research results and performance are oriented to improve computer vision and image processing integration with Unmanned Aerial System (UAS) photogrammetric data to enhance interactivity and information sharing between user and digital heritage models. The vast number of images captured from terrestrial and aerial photogrammetry will also permit to use of the Historic Building Information Modelling (HBIM) model in an eXtended Reality (XR) project developed ad-hoc, allowing different types of users (professionals, non-expert users, virtual tourists, and students) and devices (mobile phones, tablets, PCs, VR headsets) to access details and information that are not visible from the ground.

**Keywords:** Unmanned Aerial System (UAS); heritage documentation; photogrammetry; 3D modelling; eXtended Reality (XR); virtual museums; computer vision



**Citation:** Banfi, F.; Mandelli, A. Computer Vision Meets Image Processing and UAS PhotoGrammetric Data Integration: From HBIM to the eXtended Reality Project of Arco della Pace in Milan and Its Decorative Complexity. *J. Imaging* **2021**, *7*, 118. <https://doi.org/10.3390/jimaging7070118>

Academic Editors:

Giovanna Castellano, Gennaro Vessio and Fabio Bellavia

Received: 4 June 2021

Accepted: 13 July 2021

Published: 16 July 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In recent years, drone photogrammetry has become part of the daily life of many professionals in many different sectors. The main application fields are cultural heritage [1], archaeology [2], geology [3], technical and thematic cartography [4], crime and accident scene [5], human bodies survey [6], and rapid survey photogrammetry from video [7]. The starting data is always a set of photographs and therefore a set of two-dimensional digital images that are processed by the software to extract three-dimensional data. The accuracy of the work is the main advantage of photogrammetry with a drone [8–10]. Manually measuring sites can lead to several (human) errors on the part of professionals. On the other hand, drones avoid these errors and allow us to obtain accurate data, allowing surveys to be carried out with greater accuracy, significantly increasing safety at work and allowing us to reach places that are difficult for humans to access.

Thus, the Architecture, Engineering and Construction (AEC) sector is increasingly concerned with advanced simulation to create objects that behave and look as authentic as possible.

Furthermore, in the digital cultural heritage (DCH) domain, interest is growing in describing the geometry and behaviour of 3D objects through a scan-to-BIM process [11,12]. Accordingly, the integration of image processing, computer vision and 3D modelling have become significantly more useful in architecture, engineering, advanced prototyping, healthcare, and design [13–16]. In this context, the most important new movement in graphics is increasing concern for modelling objects, allowing professionals to increase the graphic value and information of heritage buildings, monuments, and archaeological sites.

For these reasons, the authors propose a scientific method based on digital photogrammetry, laser scanning and Building Information Modelling for Heritage artefact (HBIM), where it has been possible to go beyond the main 3D representation techniques, obtaining digital representations capable of communicating high Levels of Detail (LOD) and Levels of Information (LOI). On the other hand, the digitisation process of a historic building and its morphological and typological complexities requires high skill and knowledge of professional software capable of transforming simple images and 3D scans into informative models. The study here follows the workflow from the 3D photogrammetric survey phase to the digital delivery and presentation of the results through an eXtended Reality (XR) project that allows many users to employ different devices to access data and information directly accessible in other ways. Regarding this aspect, many techniques can be used to acquire the shape of ornaments and statues, but doubtless the most efficient way is using an Unmanned Aerial System (UAS). A piloted aircraft lets the surveyor get very close to the objects without wasting money hiring cranes or platforms to reach the top of the arch.

## 2. Motivation and Main Contributions

The authors' research in recent years has focused on improving the scan-to-BIM process of historic buildings, proposing and developing methods capable of automating the generative process of the model and maintaining high LODs and LOIs. On the other hand, in recent years, the construction sector has witnessed an epochal change that has led to re-engineering of the daily practices of architects, engineers, and archaeologists. Thanks to the benefits found in 3D surveying, digital photogrammetry (terrestrial and aerial) and the advent of XR development platforms, it has been possible to improve the use of scan-to-BIM models for interactive environments, increasing information sharing and reaching time for a wider audience such as students and virtual tourists. For this reason, this article proposes a method capable of enhancing the use of HBIM models to develop immersive XR environments, where the user can interact with new levels of interactivity.

The research method is based on four main research phases:

Data collection: the data collection and analysis phase envisaged primary and secondary data sources. Authors digitised the Arco della Pace monument through 3D survey techniques such as terrestrial, aerial photogrammetry and laser scanning. The main outputs of this phase are point clouds, orthophotos and mesh textured models both for the architectural elements and for the decorative apparatus composed of low reliefs and sculptures. These outputs are considered primary data sources. Secondary data sources, on the other hand, include different types of analyses and studies. The latter were conducted on various historical texts to better understand the construction technique of the monument, its historical and cultural background and the artistic values of the decorative apparatus.

Scan-to-BIM process: the digitisation process of the monument is described to show how many primary data sources have been processed sustainably, transformed into digital models capable of interacting like the latest generation applications in the construction sector. Consequently, the process of transformation and orientation of the digital models had to be based on a scientific study capable of considering different levels of interoperability of models from images, point clouds and textured mesh models.

Information mapping: once the various 3D objects have been created, a phase of mapping information is undertaken both of a visual and graphic nature (physical and mechanical characteristics of the materials, masonry stratigraphy, historical phases) and of a historical and cultural character through texts and descriptions. This phase allowed the authors to move from simple static models to objects capable of communicating different types of information.

Information sharing: finally, tests and studies were conducted to improve the models' interactivity level. Thanks to the definition of sustainable digital workflow, it has been possible to transform static models into interactive virtual objects capable of maintaining the quality of the virtual experience. Finally, the most advanced forms of virtual and

augmented reality have been tested in order to reach different types of users and the latest generation devices.

The article is structured as follows:

- a first part is dedicated to state of the art, divided in turn into a synthetic framework oriented towards HBIM and the forms of XR for the built heritage and a framework on aerial photogrammetry and its regulatory context;
- a description of the case study both from a historical-cultural point of view and from a geographical and regulatory point of view;
- the description of the method that has enabled the transformation of simple points and mesh models from 3D survey and digital photogrammetry into complex digital models (NURBS and HBIM) and XR projects with different levels of interactivity, information and immersion;
- A concluding part dedicated to a discussion of the results through a holistic approach and related conclusions.

### 3. State of the Art

#### 3.1. State of the Art about Heritage Building Information Modelling Oriented to eXtended Reality (XR)

The monument of the Arco della Pace presents numerous complexities, both from a constructive point of view and from an architectural and decorative point of view. The digitalisation process consequently required high knowledge of 3D modelling and BIM to transmit geometric, metric, and informative values at the same time. As anticipated in the previous paragraphs, the urgency of communicating information to different types of users required the study and integration of varying representation techniques, from descriptive geometry to the scan-to-BIM process [17–22]. The latter, in recent years, has shown how, through the application of specific scan-to-BIM requirements and grades of generation (GOGs) [23], it is possible to go beyond the modelling of parametric objects included in the default libraries of the main BIM applications such as Autodesk Revit and Graphisoft Archicad [24–26]. As known, such applications allow users to add information to three-dimensional objects, which in turn represent the architectural and structural components of buildings [27–29]. In the early 90s, BIM was developed for the management of new buildings, where the use of object libraries corresponding to early walls of geometric irregularities, including a wide range of choice between standard objects such as doors, windows, floors, false ceilings, and furnishings, allowed the user to faithfully represent his project and associate it with information of a physical, mechanical nature, etc. Consequently, the entire construction sector has been transformed in its daily practices, where architects and engineers have had to face a significant change, passing from the first to the second digital era. The transition almost entirely saw the abandonment of representation and manual drawing favouring CAD vector design and then subsequently BIM [30–32].

On the other hand, the benefits brought about by this new method and tools laid the foundations for the definition of new fields of applications based on digital models, such as restoration, energy analysis, finite element analysis, estimation, construction site and many others [33–36]. Consequently, the urgency of orienting the digitisation project of the existing building involved integrating 3D survey techniques with BIM. In particular, thanks to laser scanning and digital photogrammetry, it was possible to lay the appropriate foundations to represent existing buildings correctly [37–39]. The last decades have been characterised by the definition of new innovative methods, guidelines and standards that have defined this new field of application. Furthermore, interesting research in representation, geomatics, and restoration has proposed methods capable of speeding up the digitisation process of complex elements of historic buildings [21,40–43]. In this particular context, as is well known, complex vaults systems, irregular walls and decorative devices require advanced modelling techniques, where BIM modelling tools do not allow for fast and faithful representation of the building surveyed [44–46].

For this reason, the authors' research in recent years has focused on the definition of workflows capable of representing historic buildings characterised by high levels of detail and information. The definition of these methods also required an in-depth study of computer programming, through which it was possible to develop not only sustainable application methods but digital tools capable of improving the level of automation of the scan-to-BIM process [47–49]. Thanks to previous studies, the case study of the Arco della Pace monument, its architectural, cultural, historical, and artistic complexities, have laid the foundations for proposing a method capable of going beyond what is defined today in the international panorama of digital cultural heritage [50–52].

### 3.2. State of the Art about Regulation for Flying Drones in Italy and Europe

In the last decades, the use of Unmanned Aerial Vehicles in the construction and architecture field underwent significant development thanks to the increasing ease of use in piloting the vehicles and thanks to the better quality of the photographic sensors. UAVs are employed in different scenarios in the AEC (Architecture, Engineering and Construction) sector; high-resolution photographic sensors, IR sensors and thermal ones are widely used for monitoring and inspection purposes [53]. UAVs let the operators get close to the structures and buildings, preventing cranes and safeguarding the workers. Nevertheless, the time required for monitoring and inspection activities is shorter than using other terrestrial vehicles.

In the same way, the high number of images that can be acquired during a flight let the operators use the images for photogrammetric projects. In these cases, the flight had to be planned carefully to ensure the minimum overlap among the images and to avoid lack of data at the end of the elaboration. This technology permits us to reach and acquire parts of the buildings and structures that are unreachable with other instruments or require significant efforts to be mapped.

In the Cultural Heritage field, buildings are often decorated with complex friezes and ornaments that produce shaded areas when they are surveyed from ground level, both with laser scanning techniques and photogrammetry. For these reasons, in the last years, different operators decided to adopt UAVs in their daily working activities. Therefore, the regulatory bodies were forced to emanate rules to prevent accidents and interferences with regular commercial and touristic air traffic. In Italy, the first regulation concerning UAVs was enacted by the National Authority for Civil Aviation (ENAC, Ente Nazionale per l'Aviazione Civile) on the 16th of December 2013, then many editions and amendments were issued up to the present today. On the 31st of December 2020, the European regulation became effective, significantly changing many articles of the previous Italian regulation [54].

At the time of the UAV survey of Arco della Pace in Milan, November 2020, and now at the time this article is being written, the Italian and European regulations are still effective until the 1st of January 2023. The delays of the emanation of a unique and clear law are due to the COVID-19 pandemic; in fact, the transition process and alignment to the new regulation by the producers of Unmanned Aerial System (UAS) should have been completed by the 1st of January 2021. On this date, in Italy, the first edition of the UAS-IT regulation that transposes the implementing UE regulation 2019/947 concerning rules and procedures to fly Unmanned Aerial Vehicles was enacted (Table 1).

**Table 1.** List of regulations issued over the years in Italy and Europe.

Document	Edition	Date	No. of Pages
Italian Regulation (UAV)	First Edition	16/12/2013	21
	Second Edition	16/07/2015	37
	First Amendment	21/12/2015	37
	Second Amendment	22/12/2016	37
	Third Amendment	24/03/2017	37
	Fourth Amendment	21/05/2018	37
European Regulation (not effective in Italy)	First Edition	24/05/2019	27
Italian Regulation (UAV)	Third Edition	11/11/2019	37
	First Amendment	14/07/2020	37
European Regulation (effective in Italy)	First Edition	31/12/2020	27
Italian Regulation (UAS-IT)	First Edition	04/01/2021	20

This situation generated some difficulties in understanding which rules were effective at the time of the survey and how best to acquire permission to perform the 3D survey of the monument. The “No of Pages” column of the table above shows that the regulation and amendments changed over the years by adding, removing or slightly modifying the contents of single articles.

### 3.2.1. Italian Regulation UAV

Looking at the Italian regulation, it is interesting to analyse the changes between the first, second and third edition of the Italian Regulation (UAV) and the first edition of the Italian Regulation (UAS-IT). As it is clear from the title, the last Italian regulation is now aligned with the European one, focusing attention on the system: vehicle and radio control station. Now the acronym used is UAS and no more UAV. Moreover, it is interesting to notice a difference of more than 16 pages between the first and second editions of the Italian regulation. The last edition of 04 January 2021 has 17 pages less than the second edition of the 16 July 2015. The first edition (16 December 2013) of the Italian regulation is divided in 6 sections that include 26 articles. This version of the regulation is composed mainly of definitions and references to other rules and laws concerning airworthiness. Sections 2 and 3 differentiate between UAVs according to their weight, and there are different procedures for flying UAVs if they weigh more or less than 25 kg. There is general information about UAVs weighing less than 2 kg that may follow simplified procedures to get permission to fly. The procedures to acquire permission are not clear at all. In this case, general advice is provided to contact the ENAC body via e-mail and ask permission to fly by describing the pilot’s activities during the flight operations. The article regarding the pilot is unclear; in fact, it is asked that he/she holds a civil or sport flight licence and he/she must know the air rules and must have completed a training period at unclearly defined companies. Two different entities are identified in the regulation: the operator is the owner of the UAV and is responsible for the maintenance of documentation and the UAV itself, and the pilot is responsible for flight operations. A section is devoted to using UAVs for recreational purposes. Presently, there is just one article stating that UAVs can be used for recreational purposes in specified flight fields without acquiring a flight licence and without asking permission from the ENAC.

The second edition (16 July 2015) adds two sections to the previous one, namely the rules for using the air space and general rules for flying UAVs. This version of the regulation introduces some other significant changes: (i) the difference between critical and non-critical operations is addressed; (ii) the difference between specialised and non-specialised operations is defined; (iii) the procedures for flying UAVs weighing less than 2 kg are well described; and (iv) the definitions of Visual Line Of Sight (VLOS), eXtended Visual Line Of Sight (EVLOS) and Beyond Visual Line Of Sight (BVLOS) are introduced.

According to this regulation, specialised operations provide a paid service, such as video or photo recording, surveillance, environmental or industrial monitoring, agriculture services, and photogrammetry. All other activities that do not consider a payload are classified as not specialised and are considered recreational. The specialised non-critical operations are always performed in VLOS, i.e., in constant eye contact with the UAVs, far from crowds, traffic, urban areas, infrastructures, and industrial plans. In these cases, the procedures to acquire permission are much more simplified than in critical operations.

Moreover, the activities performed with UAVs weighing less than 2 kg are always considered not critical. Following the evolution of the market of consumer UAVs, and namely with the presentation of the DJI Spark, the Italian Regulation added an article regarding vehicles weighing less than 300 g. The UAVs falling in this range of weight and equipped with guard propellers did not require a flight licence or permission issued by ENAC.

In the third edition (11 November 2019), some specifications are added. In the third edition of the regulation, exams are differentiated according to the kind of operations to be conducted. In the case of non-critical operations, the certificate was issued after passing an online test, but on the other hand, to acquire the certificate for critical operations, a practical exam was needed. The article regarding UAVs weighing less than 300 g was revised: the limit is now fixed to 250 g, and at least a theoretical exam is needed to fly these lightweight UAVs. This constraint was introduced because of the large consumer market of small UAVs equipped with high-resolution camera sensors. In fact, problems started to arise linked with public security and privacy since everyone older than 18 years could have bought a UAV in a supermarket and started to fly almost everywhere without knowing the basic rules of flight. In the third edition, the minimum age to drive a UAV is lowered from 18 to 16 years. Regardless of the weight or the operation, specialised or recreational, every kind of UAV must be registered in a national online database, and a QR-code must be applied to the UAV itself. The UAV operator performs the registration that must also indicate the personal data of the pilot and its certificate. The online database that manages all the activities in the Italian airspace is [www.d-flight.it/new\\_portal](http://www.d-flight.it/new_portal) (accesed on 14 July 2021) [55].

### 3.2.2. Italian Regulation UAS-IT and European Regulation

The first edition of the new UAS-IT regulation issued on the 4 January 2021 changes the previous regulations' vision and structure completely. This last regulation receives all the indications of the European one and disciplines the aspects that rely on the competence of the state member.

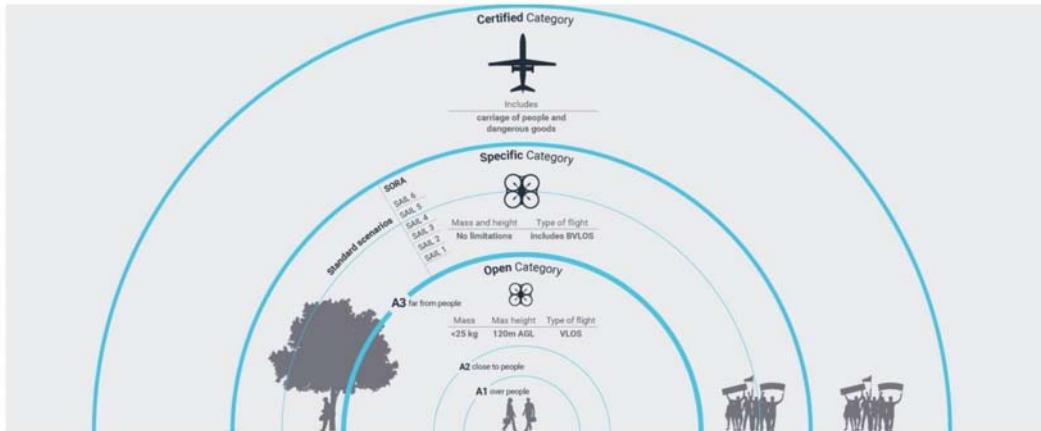
The UAS-IT regulation currently only has 5 sections, 20 pages, and 31 articles. It is shorter than the first regulation of 2013. Unfortunately, this contraction in length implies more difficulties in understanding the global view of the regulation. In fact, there are legal references to 14 different documents, (Regolamento UAS-IT, Codice della Navigazione, Regolamento (UE) n. 2018/1139 "Regolamento Basico", Regolamento (UE) n. 2019/947, Regolamento (UE) n. 2019/945) that a UAV operator should know to understand the regulations completely. In place of the weight subdivision, the concepts of: (i) Open, Specialised and Certified Categories, (ii) class identification label, (iii) Specific Assurance and Integrity Level (SAIL) are introduced.

The Open category is a category of UAS operation that, considering the risks involved, does not require prior authorisation by the competent authority nor a declaration by the UAS operator before the operation takes place. The Specific category is a category of UAS operation that, considering the risks involved, requires authorisation by the competent authority before the operation takes place, considering the mitigation measures identified in an operational risk assessment, except for specific standard scenarios where a declaration by the operator is sufficient. The Certified category is a category of UAS operation that, considering the risks involved, requires the certification of the UAS, a licensed remote

pilot and an operator approved by the competent authority to ensure an appropriate level of safety.

The Open category is itself subdivided in 3 sub-categories A1, A2, and A3, which may be summarised as follows (Figure 1):

- A1: fly over people but not over assemblies of people;
- A2: fly close to people;
- A3: fly far from people.



**Figure 1.** Categories Scheme, <https://www.easa.europa.eu/document-library/easy-access-rules/easy-access-rules-unmanned-aircraft-systems-regulation-eu> (accessed on 14 July 2021).

Each sub-category comes with its own sets of requirements. Therefore, it is important to identify which rules apply and the type of training needed in the Open category. Then, a UAV with the proper class identification label (C0, C1, C2, C3, C4) must be chosen (Table 2). Today, not even one UAV on the market has a classification label, so until the 1st of January 2023, the identification label is substituted by weight classes.

**Table 2.** Open category scheme after the 1st of January 2023, <https://www.easa.europa.eu/domains/civil-drones-rpas/open-category-civil-drones> (accessed on 14 July 2021).

UAS		OPERATION		DRONE OPERATOR/PILOT		
Class	Maximum Take Off Mass (MTOM)	Subcategory	Operational Restrictions	Drone Operator Registration	Remote Pilot Competence	Remote Pilot Minimum Age
Privately built	<250 g	A1 (can also fly in subcategory A3)	May fly over uninvolved people (should be avoided when possible) No flying over assemblies of people	No, unless camera/sensor on board and drone is not a toy	No training needed	No minimum age
C0					Read user manual	16, no minimum age if drone is a toy
C1	<900 g		No flying expected over uninvolved people (if it happens, should be minimised) No flying over assemblies of people	Yes	Read user manual Complete online training Pass online theoretical exam	16

Table 2. Cont.

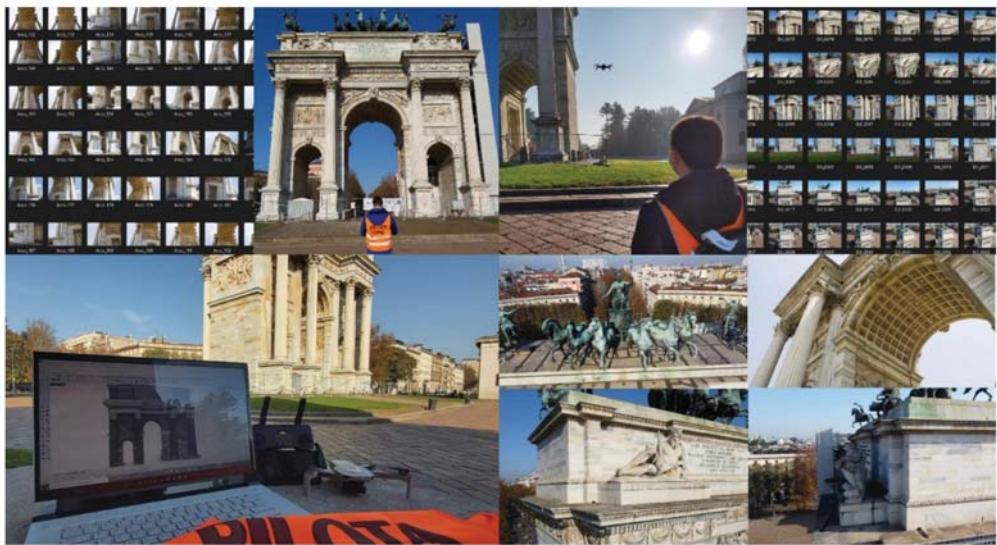
UAS		OPERATION		DRONE OPERATOR/PILOT		
Class	Maximum Take Off Mass (MTOM)	Subcategory	Operational Restrictions	Drone Operator Registration	Remote Pilot Competence	Remote Pilot Minimum Age
C2	<4 kg	A2 (can also fly in subcategory A3)	No flying over uninvolvled people Keep horizontal distance of 30 m from uninvolvled people (this can be reduced to 5 m if low speed function is activated)	Yes	Read user manual Complete online training Pass online theoretical exam Conduct and declare a self-practical training Pass a written exam at a recognised entity	16
C3	<25 kg	A3	Do not fly near people Fly outside of urban areas (150 m distance)	Yes	Read user manual Complete online training Pass online theoretical exam	16
C4						
Privately built						

If the activities do not fall under the Open category, the operator needs an operational authorisation from the National Aviation Authority. In this category, a risk assessment is needed, and there are six different levels of risk identified by roman numbers, with each level described inside the Joint Authorities for Rulemaking of Unmanned Systems (JARUS) guidelines on Specific Operations Risk Assessment (SORA). Working with UAVs is quite complex because the rules are constantly evolving, and in the last two years, the concept of regulation changed completely, passing from weight and type of operation classification to a classification based on the risk of activities. In this scenario, it was not easy to approach the survey of the Arco della Pace because it meant flying in the city centre of Milan, very close to people, traffic and inside a no-fly zone. Lastly, it must be considered that if these operations are conducted without respecting the law, the penalties are the same as the Civil Aviation Code, starting from tens of thousands of euros.

#### 4. The Research Case Study: Historical and Cultural Background, Monument Location and Flight Restrictions

##### 4.1. The Arco della Pace in Milan: Origins and History of the Arco

The Arco della Pace can be considered as the only example of a triumphal and monumental entrance to Milan, with its symbolic and commemorative presence. The arch is in the place of arrival of Corso Sempione in connection with Paris, or at the Porta Sempione, which for decades was the entrance to the city of Milan (Figure 2). The arch assumed enormous urban importance after the demolition, in 1801, of the star of the sixteenth- and seventeenth-century fortifications, when a new access road was traced and a new door was built on the axis of the Castello Sforzesco. The urban importance of the Porta del Sempione becomes substantial when one thinks that before this, the roads connecting Milan with the territory to the north-west avoided the Castle and penetrated the city, passing through openings at the points of union between the Castle and the Spanish walls called “portelli”. As a result, the place mutates in meaning: from a barrier, it becomes a passage and therefore a point of attraction towards the city centre, becoming, during the Napoleonic Empire, the main entrance to Milan.

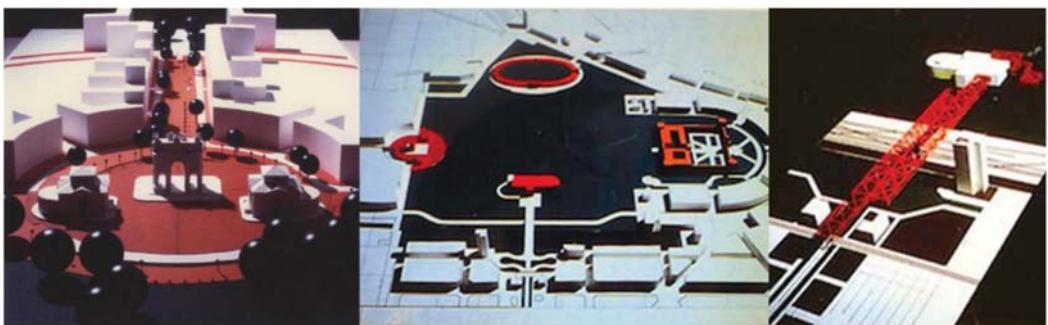


**Figure 2.** The research case study: images from terrestrial (left) and UAV survey (right).

On the 15th of May 1796, with the entry into Milan of the Napoleonic troops, and with the founding of the Cisalpine Republic first and then of the Kingdom of Italy, a positive period of effective possibilities began for the building and urban reorganisation of the city. In 1806, a “Commission of Architecture and Fine Arts” was set up, delegated to indicate the general directions for the arrangement of the city public spaces. The commission, composed of Bossi, Canonica, Appiani, Podestà, Brivio, Cagnola and Zanoja, set up an extensive urban restructuring program, with references to the French tradition of embellishment. However, innovative aspects were also introduced, both in the overall vision of the settlement and in the variation of the structure in the urban fabric. The “Plan des artistes” of 1793 and the “Plan of the embellishment” for 1798 in Paris were the reference planes.

Among the interventions suggested by the commission were an arch to be erected at the Sempione barrier and the completion of the Eastern Gate, the arrangement of the Porta Vercellina, the decoration of the Forum barracks, the decoration of the Amphitheater, the construction of a bridge between the district of S. Andrea coni Boschetti and the Collegio Elvetico, and the decoration of the Palazzo dei Giardini Pubblici.

One of the most recent enhancement plans for the Sempione Park and all its monuments is the work of Vittoriano Viganò in 1954. It was conceived as a recovery plan for a part of Milan historically homogeneous in its monumental identity. Due to the increase in road traffic and disinterest in the post-war period, it entered a state of neglect and decay. The plan conceived by Viganò is based on an idea of an urban relaunch primarily involving this large area and its identification, in the sense of open, public, and recognisable space, which can be enjoyed in various areas (Figure 3). The plan covers an area of approximately one million square meters. The Sempione system is an urban and architectural complex, rediscovering its own identity, connections, and entirety, that will come to be born as a new major attraction in Milan.



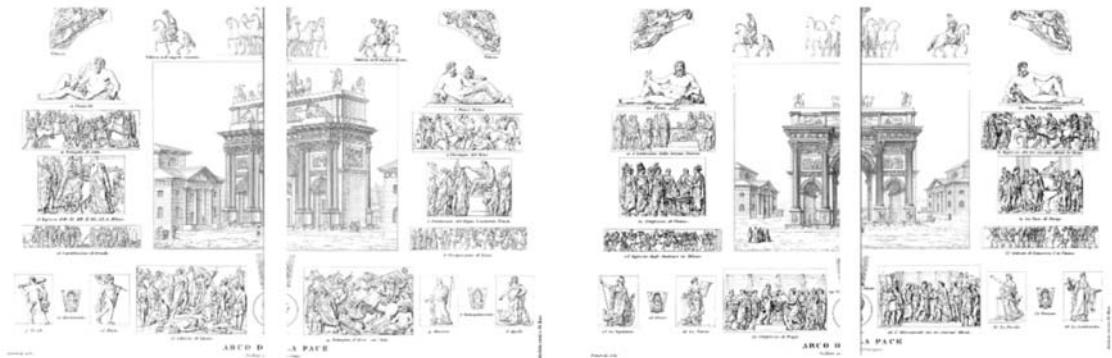
**Figure 3.** Vittoriano Vigano's plan: model of Piazza Sempione for the enhancement plan of Parco Sempione and its monuments (1996). (Location: Milan (MI), La Triennale di Milano Foundation, Photo Archive of the Milan Triennale, TRN\_XIX\_03\_0145).

Its eventual unification and modernisation could contribute to the functional revival of the monuments and transform the park heritage from an interval of the urban continuum into a homogeneous part of characteristic significance, capable of extending the historic core towards the north-west in a unique way. The start-up of the plan was set up in parts, and Piazza Sempione corresponds with what is defined as the first intervention unit (1980–1986). This intervention is followed by others that concern the whole system of the park up to Piazza Castello. Since the plan is very ambitious, it sparked various “appetites” that tried to oppose the “Franciscan force of non-speculable space, his life is difficult, and the management was slow and laborious”.

Suffice it to say that the plan, introduced in 1955, only became operational in the 1990s. The municipal administration decided to undertake a general restoration of the park and Piazza Sempione with the Arco della Pace. The intervention involved a new fence for the park, the renewal of the roads, and park botanical and floristic renewal. One of the intentions of the project was to bury all the driveways to incorporate Piazza Castello and the first stretch of Corso Sempione.

#### 4.2. Ornamental and Decorative Elements

The Arco della Pace is rich in decorative and ornamental elements, which underwent some retouching to represent the new Austrian course (Figure 4). However, the eight allegorical bas-reliefs on the pedestals were already present at the time of the new plan. The will of the central Congregation was to remind its citizens of the achievements that contributed to the Kingdom’s birth. However, since the historical events that they wanted to represent were too abundant to affect a triumphal arch, only those relating to the most important events were chosen. Allegorical and allusive figures were then chosen to evoke “the beautiful arts, the fertility of the Lombard soil, the historical events” that were most significant. To these are added other events, including the Congress of Prague, the Meeting of the Three Great Allies and other war enterprises, thanks to which the much-desired peace was obtained. Passage of the Rhine, Capitulation of Dresden, Battle of Ar-cis-sur-Aube, Occupation of Lyon, the Battle of Paris and, finally, the triumphal entry of the three monarchs into the city of the French Empire. These findings were the works of many sculptors including Camillo Pacetti, Luigi Acquisti, and Pompeo Marchesi. In addition to the military enterprises, the political operations that made the Peace of Paris and the Congress of Vienna official were also mentioned.



**Figure 4.** Historical reports: the ornaments of the Arco della Pace. Overall table of the elements relating to the front of the monument towards the Castello Sforzesco (**left**) and Corso Sempione (**right**), taken from the publication edited by G. Reina and published in 1856. Di Baio historic archive. (Giani, G., L'Arco della Pace di Milano, Di Baio Editore, Milano, 1988).

#### 4.3. Monument Location and Flight Restrictions

As anticipated, the monument is in the city centre of Milan in the centre of Piazza Sempione, an important city hub (Figure 5).



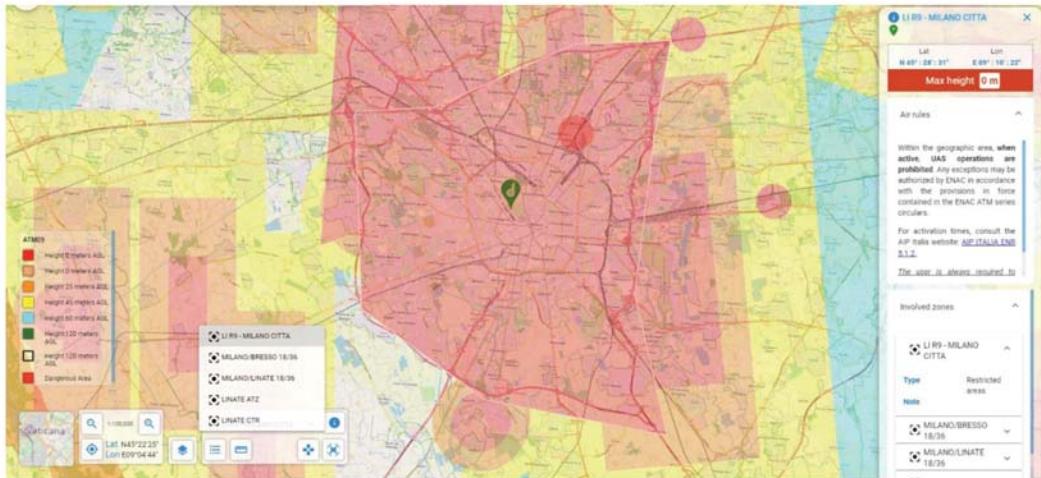
**Figure 5.** Google Map image centred on the monument.

Here ends Parco Sempione, one of the biggest parks of Milan. Local people and tourists always crowd this place; moreover, Piazza Sempione is half surrounded by jammed streets and punctuated by dehors of the adjacent bars and restaurants.

Considering the regulations, both Italian and European, flight in such places is forbidden because it will interfere with public activities around the monument. Moreover, Italian regulations on airspaces subdivide the national ground into different zones with special rules concerning flying with manned or unmanned vehicles. Consulting the aeronautical

maps provided by the Italian online database for UAVs, it appears that the area of the survey falls into four different restriction areas where flight is forbidden to anyone, mainly for security reasons (Figure 6):

- LI-R9 Milano-Città;
- Milano/Bresso 18/36;
- Milano/Linate 18/36;
- Linate Aerodrome Traffic Zone (ATZ);
- Linate Control Traffic Region (CTR).



**Figure 6.** D-Flight Map image centred on the monument, with restriction areas highlighted. Each colour refers to a height limit for flight above the city of Milan. In the red areas, flight is prohibited. Source: [www.d-flight.it/newportal](http://www.d-flight.it/newportal), accessed on 14 July 2021).

The city authorities can obtain temporary permission by submitting all the necessary documents and a high detailed relation that describes the activity, the timetable of the flights, the risk assessment and the precautions taken to decrease the level of the risk. The authors, both holding a piloting license, provided the material mentioned above to the prefecture of Milan that has the faculty to issue the permission, then permission also had to be approved by the ENAC. Even if the prefecture issues permission, the Authority can revoke it. A month after submitting the request, the survey activities described in relation received a positive judgment from the two authorities. The pilots considered a sufficient buffer area around the monument, and they chose a date that fell in the lockdown period linked with the COVID-19 pandemic. Therefore, all the shops' restoration activities were closed, and there were no crowds around the monument due to the prohibition on staying in public spaces without a proven reason.

## 5. Material and Methods: From Geometrical Surveys to HBIM, Virtual Museums and eXtended Reality

The method proposed in this paragraph has been structured in an attempt to outline an operational workflow that is as sustainable as possible (Figure 7). The key factors for improving the scan-to-BIM-to-XR process of the monument were:

- Integration of aerial photogrammetry in the building digitisation process to complete the textured digital model;
- 3D mapping able to be automatically recognised through the real-time synchronisation of multiple environments, from NURBS modelling software and BIM platforms to XR development platforms;

- Interoperability and synchronisation of digital models in various environments; automatic recognition and real-time synchronisation of digital models through the main 3D exchange formats (open source and not) such as the 3DM, DWG, RVT, FBX, OBJ;
- The interactivity of XR projects; through IT development based on VPLs and Blueprints, it has been possible to create interactive virtual objects capable of interacting with all user inputs on different kinds of devices (tablets, laptops, PCs, and mobile phones).

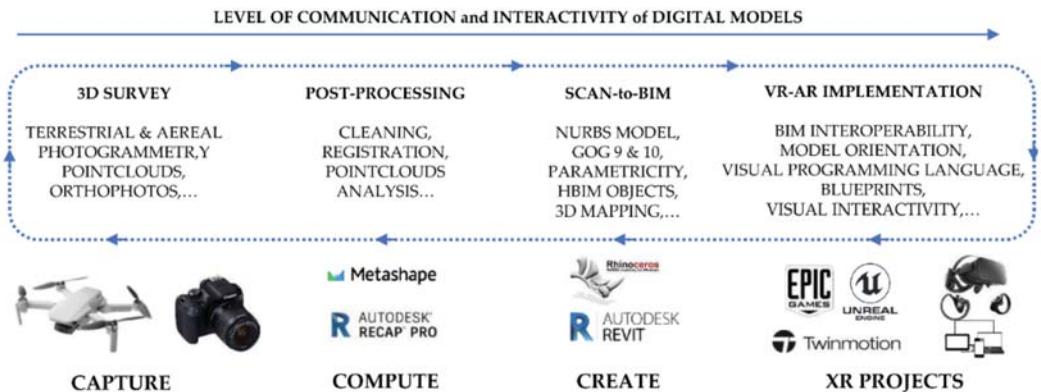


Figure 7. The digital workflow applied to the research case study.

### 5.1. UAV Photogrammetric Survey

Since the Italian regulation about airspaces and the upcoming European law consider special rules for UAVs weighing less than 250 g, it was decided to use a UAV with this peculiar characteristic. Moreover, the Italian law obliges the drone pilot to install guard propellers to use the drone in a public space, such as Piazza Sempione, where the monument is. For these reasons, the survey team employed a DJI Mavic Mini for the photogrammetric survey (Table 3). The photogrammetric survey was sided by topographic measurements performed with the Leica TS12 (Leica Geosystems AG, Heerbrugg, Switzerland). A simple network of four points was created around the monument, two vertices are linked through the open passing arch in the centre. From the station points, some natural points were measured on three sides of the monument (one side is covered from the base to the top with scaffolding for restoration activities) to check the correctness of the photogrammetric elaborations. The aim of the survey was to collect enough data to represent the arch by means of a 3D model at a 1:50 drawing scale.

Table 3. Specification of DJI Mavic Mini.

DJI Mavic Mini—Specs	
Sensor size (pixel)	4000 × 3000
Sensor size (mm)	6.48 × 4.86
Pixel size (mm)	0.00162
Focal length (mm)	4.49
Flight time (min)	28

Considering the DJI Mavic Mini specification listed above, it is possible to calculate the mean distance of acquisition to get the desired resolution. It was decided to assume the “plotting error” (p.e.) as the parameter to calculate the distance of acquisition. This value derives from the cartography field and is related to the precision of a map, and it is conventionally assumed to be equal to 0.2 mm. The p.e. obviously changes in relation to the scale of the map; 0.2 mm must be multiplied by the scale factor. Consequently, the p.e.

at a 1:50 scale is equal to 1 cm. Conventionally in cartography, the sampling measurements tolerance is assumed equal to 2 times the p.e. Moving from the 2D cartography field to the photogrammetric 3D domain, it was decided to impose a Ground Sampling Distance (GSD) equal to the p.e. at 1:50.

$$p.e.1:50 = 0.2 \text{ mm} \times 50 = 1 \text{ cm} = GSD_{1:50} \quad (1)$$

The distance of acquisition to reach at least this value at the end of the survey is computed with the equation:

$$c:D = p_x:GSD$$

$$4.49 \text{ mm}:D = 0.00162 \text{ mm}:10 \text{ mm} \quad (2)$$

$$D = 27 \text{ m}$$

where:  $c$  = focal length,  $D$  = distance of acquisition,  $p_x$  = pixel size, GSD = Ground Sampling Distance.

Consequently, 27 m represents the theoretical value considering an ideal condition with the camera placed on a tripod without external interferences. The practical activity suggests, even in good conditions, halving the distance from the surveyed object to avoid poor data at the end of the survey, and in this case, the authors decided to fly at a mean distance of 10 m from the monument. Therefore, the number of images increased significantly. From this, 945 images at 12 MPixels, the maximum resolution of the camera, were collected in JPG format during three flights to cover all the facades of the monument, the decorative apparatus and the statues that crown the top of the arch. The flights were performed in manual mode, frequently changing the orientation of the camera gimbal to capture the complexity of the shapes from different views and to cover all the possible shadow areas on the monument. As much as possible, the flights followed regular paths, performing vertical strips all around the building. The vertical (longitudinal) overlap of the images was controlled by setting the auto interval acquisition of the images equal to 2 s, and the side (transversal) overlap was valued directly on the screen by the video operator of the drone.

### 5.2. Terrestrial Photogrammetric Survey

Due to the scaffolding from the base to the top on the east side of the monument, it was decided to merge the data of the photogrammetric flight with the photogrammetric terrestrial data acquired in 2019 when the scaffolding was not in place.

That survey was performed to produce a parametric model with Rhinoceros, starting from the dense point clouds computed at the end of the photogrammetric process. The same procedure was adopted on that occasion, and the photogrammetric survey was followed by a topographic campaign. The camera used was a Canon EOS 1100D (Canon Inc. Ōta, Tokyo, Japan) coupled with an 18 mm lens (Table 4).

**Table 4.** Specification of Canon EOS 1100D.

Canon EOS 1100D—Specs	
Sensor size (pixel)	4272 × 2848
Sensor size (mm)	22.2 × 14.7
Pixel size (mm)	0.00534
Focal length (mm)	18

The distance of acquisition was calculated as before to produce a dense point cloud with an accuracy of 1 cm.

### 5.3. UAV Data Elaboration

The UAV data were elaborated following a pyramidal schema, from general to particular, using Agisoft Metashape Pro (version 1.7.2 build 12070). Firstly, all the 945 images were imported into the software; the GPS information, namely the position of the drone

during the acquisition of each single image, and the orientation of the camera were removed before the alignment phase. The elaboration considered two phases: the first was useful to elaborate all the images simultaneously and find the 3D dense point cloud of the top architectural elements of the arch, which were not visible from the terrestrial photogrammetric survey. The second phase considered only the decorations, the statues, and the bass reliefs. Starting from the results of the first alignment, the bounding box was then limited around each decorative element, and 3D mesh reconstruction was performed to access their high-resolution models separately from the architecture.

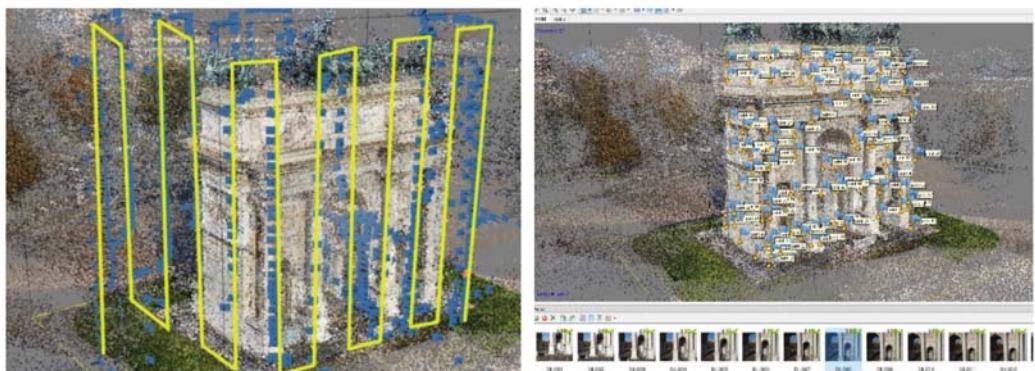
### 5.3.1. The Building

Immediately, some problems arose. In fact, even if the software said that all the images were “correctly” aligned, it appeared clear that there were some errors in the geometrical reconstruction of the building. The east side, the one with the scaffolding, was misplaced, turning by 90 degrees in the plan and giving to the arch an “L” shape. The result was always the same, even when the accuracy of the alignment was set to the highest value. Additionally, the overlap on that side seemed to be correct. Looking carefully at the images of the east and north sides, the maxi-screens on the scaffoldings were broadcasting the same images with the same timing both on the north and east edges of the arch (Figure 8).



**Figure 8.** On the left, the misalignment of the images; on the right, the maxi-screen that prevented the correct alignment of the images.

The areas of the photos with the screens were masked directly in the software, and the alignment then gave proper results. Then, the natural points were checked and placed on the images, and the topographic measurements were imported into the project (Figure 9). After the optimisation process, the mean error on the points measured by the operator is equal to 2 cm, so the model can be used to sample measurements compatible with the tolerance of 1:50 drawing scale (Figure 10). As described in paragraph 5.6, the 3D model of the arch was developed starting from point clouds data in the 3D modelling software McNeel Rhinoceros version 7.



**Figure 9.** On the left, the regular manual path followed during the flights; on the right, the natural measured points.



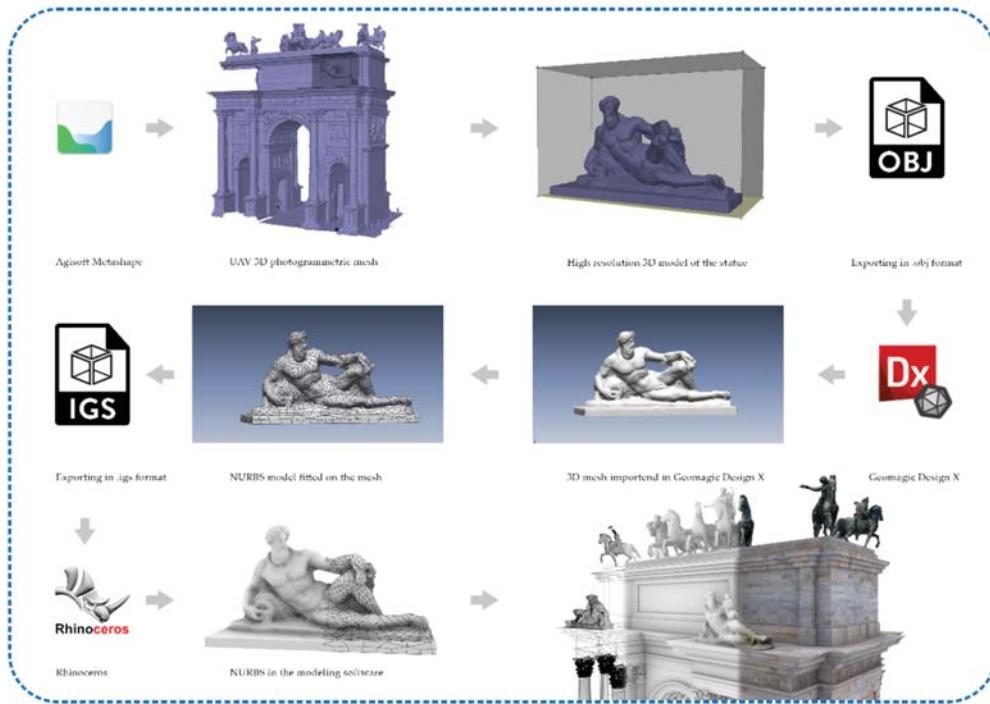
**Figure 10.** UAV photogrammetric dense cloud of the upper part of the arch.

### 5.3.2. The Statues, Ornaments, and Bass Reliefs

Elaboration of the statues, ornaments and bass reliefs followed the same pipeline of the arch, but the models were computed separately for each decorative element. This elaboration phase aims to obtain the NURBS models to be included in the general model of the arch. Unlike the classical architectonical elements, such as walls, pillars, columns, and friezes that simple geometries can describe, the statues and decorations require a different approach to generate the NURBS models. The classical elements are modelled, extracting sections and elevations directly from the point clouds, and the study of the geometries is supported by historical and design drawings.

On the other hand, it is impossible to adopt the same approach for completely free forms elements such as statues. Obviously, these elements could not be neglected in the restitution phase of the model, but it was not possible to shape them by extracting generating features and patching lines. The solution adopted was to perform the transformation with reverse engineering software, such as Geomagic Design X v 5.1. This step requires correcting typical mesh errors: auto intersecting, non-manifold, crossing, redundant, tangled, reversed faces, small tunnels, and duplicated vertices. They would cause bad results

after the auto surfacing command that fits NURBS on the targeted meshes. For this reason, the elaboration of these last elements considered the following steps (Figure 11):



**Figure 11.** The workflow to generate the NURBS geometries of the statues and the decorations.

#### Agisoft Metashape

- Duplicating the original UAV chunk;
- Resizing of the bounding box around each statue and bass relief;
- Elaboration of the depth maps at the highest resolution;
- Generation of the meshes using as source the depth maps;
- Export of the meshes in .obj format.

#### Geomagic Design X

- Import the .obj files;
- Fixing the topological errors of the meshes;
- Creating a watertight mesh;
- Auto fitting the NURBS geometries on the meshes;
- Export the meshes in .igs format.

#### McNeel Rhinoceros

- Import the .igs file without scaling or moving the single object.

#### 5.4. Terrestrial Data Elaboration

The terrestrial dataset comprises 229 images, and the elaboration phase followed the same pipeline of the UAV photogrammetric project, giving results in terms of accuracy comparable to those described above. This elaboration aimed to produce a dense point cloud of the lower part of the architecture to be merged with the dense point cloud and meshes coming from the UAV dataset elaboration (Figure 12).



**Figure 12.** North elevation of the arch, terrestrial photogrammetric point cloud.

##### 5.5. Data Merging

The two datasets did not share the coordinates because each one has its own local reference system measured with the total station. To place the two models in the same position, some manual points of the same architectonical elements were collected on both projects. The points are well distributed on three of the four elevations. Then, a new Agisoft Metashape project was created to append the terrestrial and UAV chunks. The former one was aligned to the latter using the architectonic points and fixing the scale of both models. After cleaning the overlapping parts, the two models were merged, saving the best geometries of each one, i.e., the bottom part of the terrestrial survey and the upper part of the UAV survey (Figure 13).



**Figure 13.** On the left, final merged 3D point cloud. In the centre, decorative apparatus recorded from the terrestrial photogrammetric survey; on the right, 3D model of one of the statues on the top of the arch acquired with the UAV survey.

##### 5.6. HBIM Generation: From Mesh-Textured Models to NURBS Models and Heritage Building Information Modelling

Thanks to the integration of primary and secondary data sources, the digitisation process of the monument was able to benefit from point clouds coming from aerial photogrammetry and many documentations and historical drawings capable of communicating the constructive logic of the building. As anticipated in paragraph 5.3, thanks to the inte-

grated use of point clouds and textured mesh models, it was possible to lay the appropriate foundations for defining a method capable of representing any type of shape in BIM logic. In particular, the use of GOG 9 and 10 made it possible to extract geometric primitives, slices, and wireframe models directly from point clouds and mesh models from aerial and terrestrial photogrammetry.

Figure 14 shows the multi-step approach, moving from simple points in space or mesh polygons to a NURBS model capable of corresponding to the surveyed reality. Thanks to NURBS modelling, it has been possible to create mathematical models capable of going beyond the limits imposed by BIM applications, which are still characterised by a very limited number of 2D representation and 3D modelling tools.

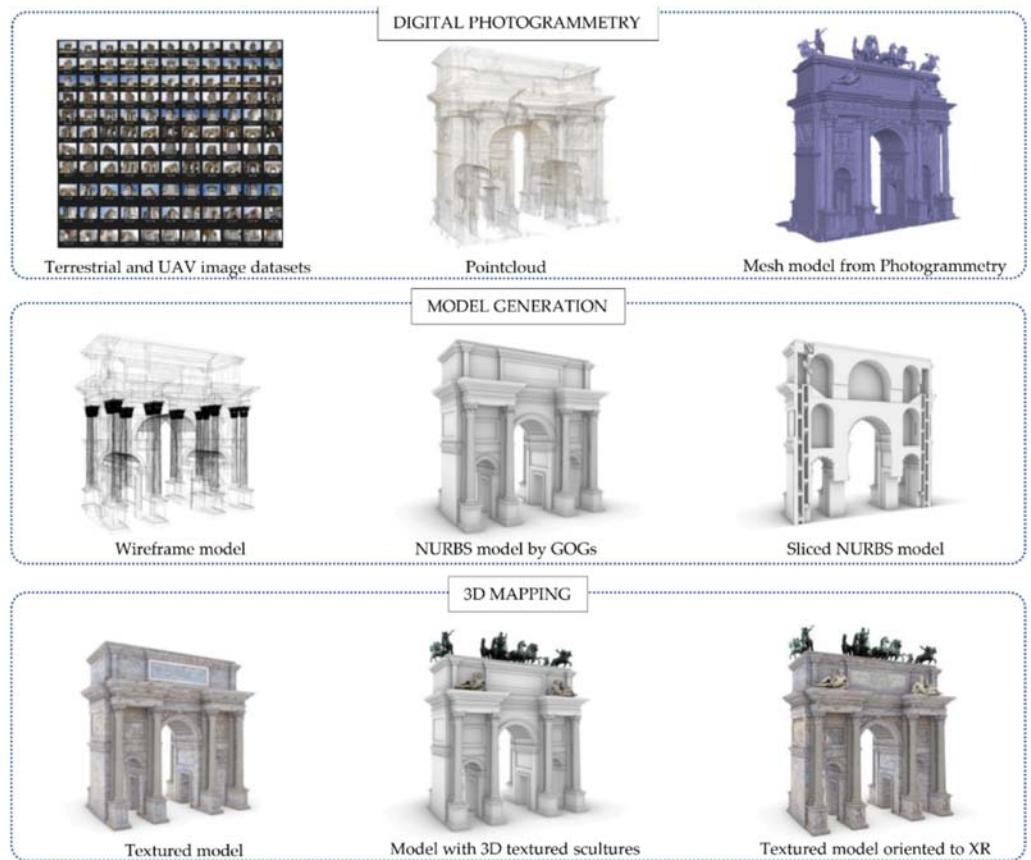


Figure 14. The scan-to-NURBS process applied to the Arco della Pace in Milan.

The second step allowed the transformation of NURBS models into HBIM objects capable of communicating high levels of information (Figure 15). Unlike newly constructed buildings, historic buildings require a 3D mapping phase capable of communicating their unique architectural, structural, material, and decorative features. Each historic building is a world unto itself, where generic materials and textures enucleated in the BIM libraries do not allow an appropriate representation of the structure, and consequently the mapping and sharing of information is not always truthful. For this reason, aerial photogrammetry has enabled a 3D mapping phase capable of communicating the integrity of the materials for each individual digitised element, from the coffered vaults, to the

pillars, up to the sculptures and low reliefs. The latter also required the transmissibility of material information and the development of new BIM parameters able to tell the story represented through textual descriptions. On the other hand, BIM applications are not easily usable by non-expert users in 3D digitalisation. For this reason, the method envisaged the development of advanced XR environments capable of reaching a wider audience and consequently enhancing the communication levels of digital models for students, virtual tourists, and other forms of users.

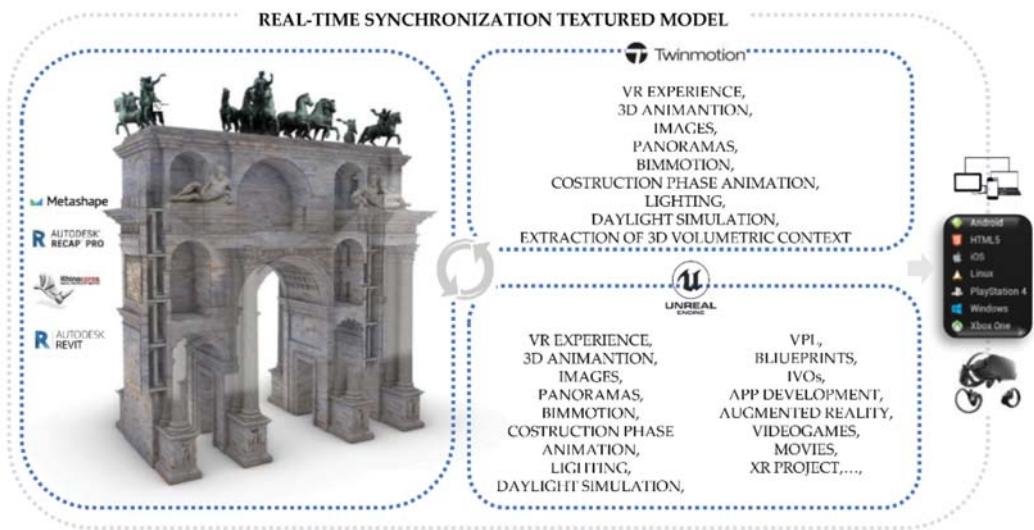


**Figure 15.** The HBIM objects of the research case study and the main 2D drawings extracted from the HBIM project browser.

Once the various NURBS objects were transformed into BIM parametric objects, thanks to the verification of the grade of accuracy (GOA), it was possible to communicate the reliability of each element created thanks to point set deviation analysis. In particular, thanks to an automatic verification system (AVS), the standard deviation between point clouds and BIM objects was calculated [23,56,57]. The value reached for every single element allowed us to define a GOA of about 1–2 mm. Consequently, the development of HBIM parameters capable of communicating this value within the propriety window of each object has allowed the user to identify the GOA, the LOD obtained, the scale of representation and the creation of schedules and databases able to accurately compute numerical quantities such as area and volume and subsequently define the materials and restoration phases useful for the conservation of the monument over time.

### 5.7. Synchronising HBIM Models with XR Development Platforms: The Virtual Visual Storytelling of the Arco della Pace in Milan

The level of interactivity achieved was, however, limited to a small circle of experts. Consequently, the passage from the information mapping phase to the information sharing phase led the authors to develop new interactivity levels, exploring the latest generation techniques and tools and defining a development process capable of immersing any type of user in XR environments (Figure 16).

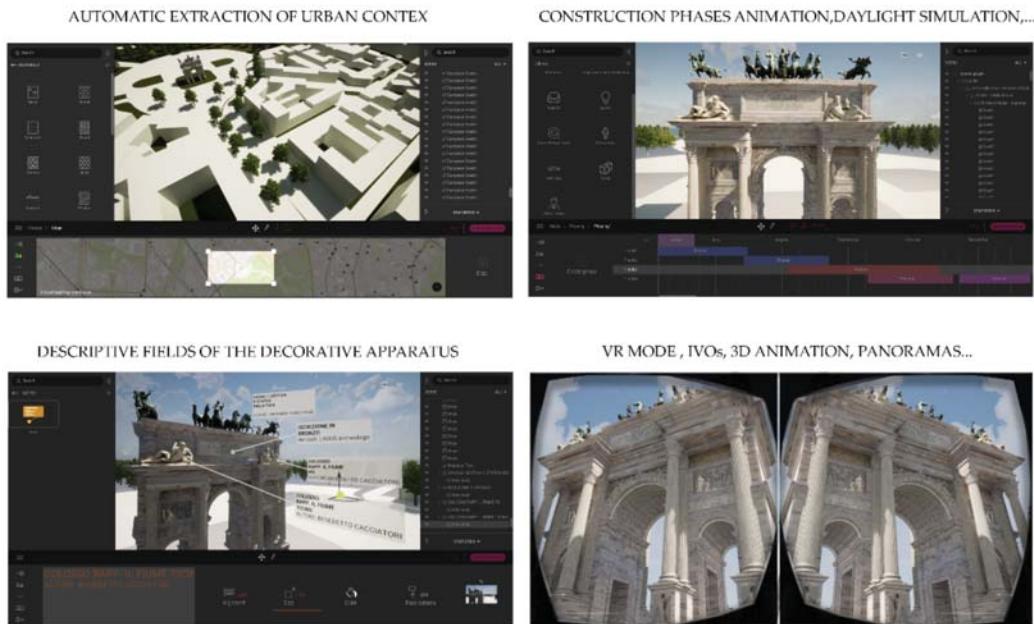


**Figure 16.** XR development approach applied to the research case study.

In this specific context, the study and understanding of the development techniques of XR environments made it possible to test and define a process based on the use of open-source platforms such as Unreal Engine, unity and Twinmotion. Unlike the scan-to-BIM-to-XR methods already consolidated in recent years, the main added value of the method was found thanks to the possibility of the synchronisation of multiple modelling software and XR development platforms, avoiding interruptions and development discontinuity between opening and closing one software to another. Thanks to the development of new add-ins, functionalities integrated into software architecture, in addition to exponentially reducing XR development times, it has been possible to define a workflow that can also be applied to experts in the construction and virtual museum sector [58], who do not always possess computer skills capable of increasing the level of interactivity of their digital models.

A second benefit of the proposed method derives from a synchronised mapping technique between Autodesk Revit, McNeel Rhinoceros and XR software such as Twinmotion and Unreal engine. In particular, the method required the use of many images from aerial photogrammetry. It has been found that the main 3D mapping techniques in modelling software and BIM platforms involve the use of decals. Decals are non-repeating textures that are applied to the surface of an object with a given projection. Decals are textures placed directly on the specified area of one or more objects. Decals are used to change a limited part of an object colour. Decals should be thought of as a single specific texture, rather than side-by-side textures, as they are when used in a material. This is an easy way to apply single images or similar textures to objects without going through complex texture mapping operations.

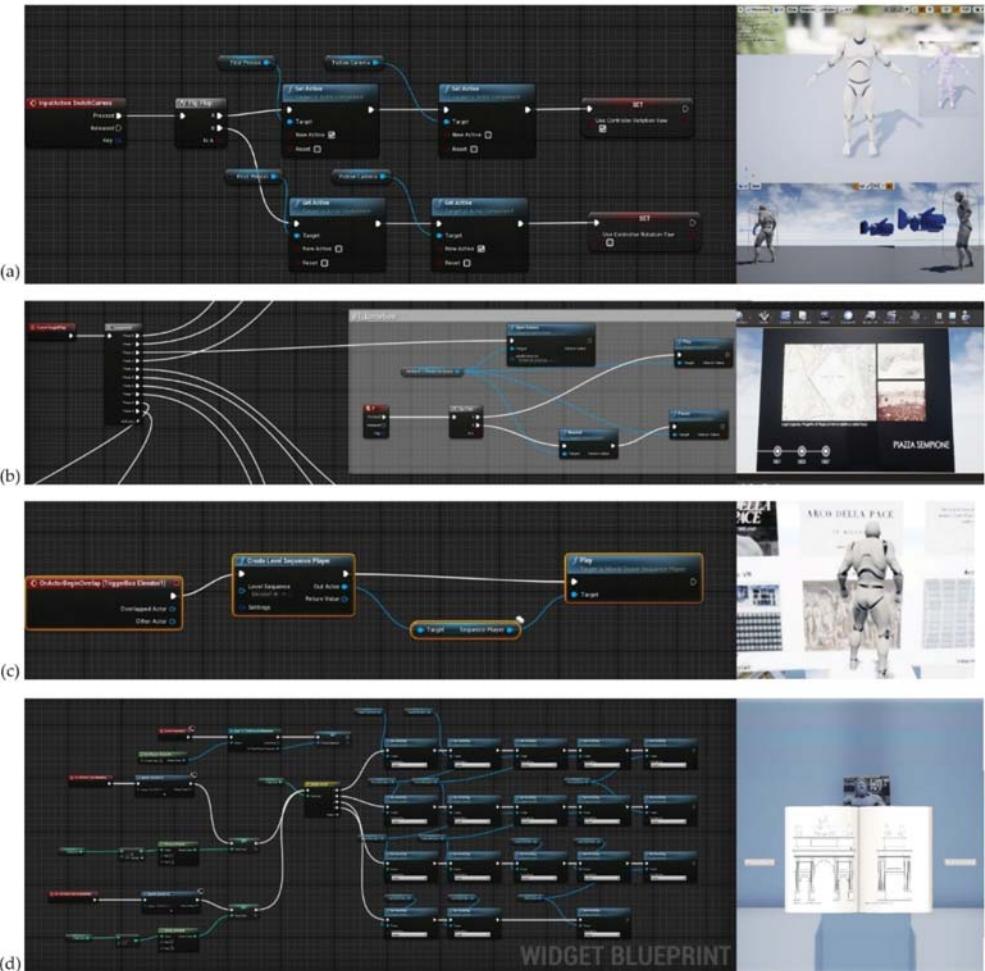
On the other hand, it was found that the mode of synchronisation of digital models with XR platforms could not include this technique capable of correctly representing the uniqueness of the monument. Furthermore, thanks to the exhibition developments in computer graphics and the development of the Twinmotion software, it was possible to reduce the phase of graphic post-production and 3D mapping, directly using textures pre-processed in NURBS modelling software and BIM platforms such as McNeel Rhinoceros, Autodesk Revit and Graphisoft Archicad. It has been found that Twinmotion is a Real-Time Rendering software for Architecture, which has recently reached its best performing version; in fact, with just a few clicks, the last version (2021) allows you to connect with the major CAD software existing on the market, such as Archicad, Revit, Sketchup Pro and Rhinoceros; thanks to Twinmotion 2021, it was possible to animate a large number of interactive virtual objects (IVOs), improving and optimising the textures; accelerating the rendering process, which is reduced to a few seconds; and using objects present in the software library or imported from the web (Figure 17).



**Figure 17.** VR project developed in Twinmotion: the development was based on the real-time synchronisation between NURBS modelling software and multiple BIM platforms (McNeel Rhinoceros and Autodesk Revit). It allows the user to obtain different interactive tools, such as the automatic extraction of the urban context, descriptive fields of the decorative apparatus, 3D animation, VR mode, construction phase animation and other types of output and interaction.

For these reasons, the 3D mapping phase had to rely on the definitions of specific textures and the reworking of graphic parameters associated with every single element created. Assuming that a texture can be applied to the surface of a 3D model to add colour, a coating or other details such as gloss, reflectivity or transparency, the problem of representing a texture in 3D rendering can be solved by UV mapping. U and V are the texture coordinates corresponding to X and Y. Think of U as the direction that goes from one side to the other of a quadrangular sheet. Think of V as the other direction, the one that goes from top to bottom. UV texture mapping is used whenever you apply an image to a material and then apply that material to a model. Texture mapping properties manage texture map projections for selected surfaces, polysurfaces, and meshes, representing a 2D image on a 3D model. The mapping transforms a 2D source image into an image

buffer called texture. Finally, the last phase of the process involved enhancing the levels of interactivity achieved in the implementation phase. Figure 18 shows the main blueprints developed for the Arco della Pace XR project.



**Figure 18.** The main Blueprints developed for the XR project: (a) change of XR mode between first and third person, (b) animation 3D, (c) interactive elements for vertical translation, (d) interactive book consultation.

The IT implementation phase envisaged specific blueprints such as Level Blueprints and Blueprint Classes, Blueprint Macros and Blueprint Interfaces. These blueprints contain the scripts necessary for the game level to react to the player's input with objects that animate, emit sounds, and change their composition based on the player's actions. The structure is also designed so that the creator of levels can reuse the same Blueprints for the same or slightly different functions, without having to redo the same work for each game element.

One type of Blueprint class is Construction Script, which kicks into action when an actor is set in the game level or updated. It serves to manage the changes that the actor needs when certain events occur. The Blueprint Visual Scripting system in Unreal Engine

is a complete gameplay scripting system based on the concept of using a node-based interface to create gameplay elements from within Unreal Editor. Blueprints is the visual scripting system inside Unreal Engine 4 and is a fast way to start prototyping your game. Instead of writing code line by line, you do everything visually: drag and drop nodes, set their properties in a UI, and drag wires to connect Object-Oriented (OO) classes or objects in the engine as with many common scripting languages. This system is highly flexible and powerful as it allows designers to use virtually the full range of concepts and tools generally only available to programmers.

In addition, Blueprint-specific markup, available in Unreal Engine C++ implementation, enables programmers to create baseline systems that designers can extend.

The latter, once developed, made it possible to create static objects without any level of interaction coming from the scan-to-BIM process described here, passing from simple static meshes to IVOs capable of responding to user input. Moreover, thanks to a process of defining the virtual-visual story telling of the monument (VVS) it was possible to tell, as well as with high levels of interactivity, even with virtual rooms where every single decorative apparatus, sculpture and low relief has been digitised and inserted in a museum itinerary. Thus, with any type of device (mobile, tablet, VR headset or PC), the user can remotely explore the intangible values reported in the XR project and become aware of the historical, cultural background of the monument and Milan.

The VVS was developed based on the following XR environments: two interactive menus, in which, thanks to the implementation of specific trigger boxes and blueprints, it was possible to migrate the user in first or third person to new levels. The bass relief, for example, are initially placed at a distance from their true position; when you approach the model, they move until they reach their respective position on the Arc. The Trigger Box is one of the actors that can be activated and cause events in the level; they are used to trigger events in response to interaction with them within the level. The Trigger Box is a trigger that can be placed in the project by dragging it into the layer. In the project, the event that activates the Trigger Box is the overlap of the avatar with the trigger. The various levels identified made it possible to define a progressive narration of the monument: from general information and multimedia files that describe the city of Milan, the square, and the cultural, historical, and geographical context of the monument up to a second menu where rooms were created dedicated to a museum display of the decorative apparatus. After placing the Trigger Boxes, the nodes are developed within the Blueprint level. The second interactive element, the video file, is played on a static mesh with the media source asset file.

The steps carried out were the following:

- creation of the Movies folder within the Content in which to place the video in .mp4;
- through File Media Source and Media Player, the video is associated with the project within the Content. The video resource is generated accordingly;
- creation of the Mesh, that is, the surface on which the video will be visible;
- Simply by dragging the video asset onto the mesh, you relate the video to the surface;
- development of nodes within the Blueprint level so that the video is played on the mesh starting from the start of the virtual experience. This happens automatically, but only through a keyboard command, “P”, which allows you to start and pause the multimedia content.

Inside some ideal rooms, books have been inserted, referring to the bibliography essential to the project. These were made as Widgets, again based on Blueprint. The books are then displayed only after pressing a key, so you can choose independently whether to display them and when. Once opened, the manual is displayed in the foreground. Within the Designer Mode, the actual book was created, inserting one page at a time and the respective drawings. Subsequently, by switching to Graph editing Mode, the nodes necessary for the animation of the book were created. One part concerns the activation of the Widget, through a button, the respective link to the avatar, and finally the sounds of the pages when they are browsed. The other part instead concerns how and when the pages

are visible. Everything is based on links between the pages so that the one you are on is the only one visible while the others remain invisible. Therefore, everything is based on a system of switches that allow these settings to be changed once the buttons next to the book have been pressed to move to the next or previous page.

Consequently, the user can immerse himself and understand the story represented through IVOs, digital archives, interactive books and multimedia files that have different kinds of content. For most of the sections, the method of creating the project based on the “Blueprint third-person template” makes it possible to change it in person thanks to creating a Blueprint that allows a quick exchange to take place by pressing a button. In this way, it is possible to better view the various sections by passing from a more immersive view, such as the first-person view, to one that allows you to better compare and understand the dimensions in relation to the height of the avatar. This was carried out within the avatar’s Blueprint, in which the second camera was set at face height to view the scene in first person.

The project is divided into levels: in the main one, there is the model of the arch, with the interactive menu to the left and right of the latter. The other levels house the in-depth rooms, which can be reached from the main room (Figure 19). The following Place Actors have been placed inside each one: AtmosphericFog, BP\_Sky\_Sphere, and ExponentialHeightFog, which complete the preliminary light setting of the levels. To import the files that make up the model, the settings that regulate the collision for each solid object can be changed so that the solids and voids are correctly processed.



**Figure 19.** The main section of the Virtual-Visual Story telling of the monument: from virtual museum to interactive virtual objects (IVOs).

Finally, the complete version of the XR project was geared towards the integrated use of VR, the Oculus Ref., a virtual reality device that allows high-quality immersive vision. It consists of a viewer, audio headphones, sensors and two controllers. The sensors are used to track the user’s movements, while the controllers allow you to interface with the experience in a more interactive way and manipulate the objects within the VR project using your hands in a rather realistic way. Through sensors, tracking allows the user to look around in the virtual environment exactly as if it were in the real world. This system allows for the most natural interaction possible, improving the sensation of immersion. Within the Unreal Engine program, it is possible to convert the project and make it compatible with vision through Oculus Rift. To do this, it is necessary to change the display settings by selecting VR Preview, expanding the menu next to the Play icon, and thus starting playback with the viewer and controller. At the end of the elaborations, the models of the 8 statues

(3 lying men, 4 knights and a chariot with 6 horses) were placed in their correct position. It was not necessary to move or scale them because we were careful not to change the reference system during each import and export phase.

### 5.8. From HBIM Models and IVOs to Augmented Reality

The proposed method has made it possible to create an immersive environment in all respects. The final user can interact with many IVOs and discover information, from historical-cultural content to precise information such as the descriptions of each low relief, sculpture, etc. Thanks to the in-depth historical research that allowed for the implementation of the VR project and the virtual museum of the monument itself, a further implementation phase was conducted with the ultimate goal of achieving an alternative form of human-computer interaction.

In recent years, several studies have developed applications capable of creating, setting up and sharing AR objects. It has been found that unlike VR, which reproduces the real world to create digital spaces, AR understands and includes the real world, superimposing virtual images on real environments, spaces and images.

In particular, AR was considered by the authors to be a suitable solution for several reasons:

- use of IVOs and HBIM objects for different purposes concerning VR,
- addition of new levels of information, in real-time and with a high rate of interaction using mobile devices of any kind, including wearable technologies,
- superimposition of multimedia information on what you are watching on any display (text, images, live or animated films),
- access to an AR system via the web through devices equipped with GPS, a web camera and an internet connection,
- use and accessibility is within reach of any type of user (expert, professional, students, virtual tourists and on-site tourists) through web apps that can be easily downloaded via the app store,
- creation of a personal account that can be implemented over time,
- ability to view objects and their information in a targeted manner, avoiding having to access the general model of the arch and discriminate between other objects,
- avoid the installation of particular software applications and use expensive digital devices,
- sharing of the model through simple links.

Figure 20 shows the web-based AR library developed for the Arco della Pace. Each object is easily navigable and viewable in AR mode. The associated information has been selected to reach the user in a targeted manner (on-site or remotely) with precise and concise descriptive texts, thus providing a cognitive approach to the decorative schema of the monument without great effort.



**Figure 20.** AR implementation: the web-based AR platform of the Arco della Pace, Milan, Italy.

### 5.9. Critical Analysis of the Proposed Workflow: Pros and Cons Found during the Implementation Process

The method proposes a continuum that, starting from the real world, leads to a completely virtual interactive world, representing “possible worlds” to create a “sense of presence and interaction” in the user. The relationship between IVOs, information and the user thus becomes the first factor of scientific investigation. The established relationships must be deeply tested in this specific context, guaranteeing the best possible experience from different perspectives. Theoretical assumptions on the use of virtual realities have been dealt with in-depth by the technologist Giti Javidi [59], who has identified positive theoretical correlations between constructivism and virtual learning environments. Through these developments, it has been suggested that, using the XR project, hundreds of specific objectives can be pursued by different means (texts, discussions, videos, software, podcasts, etc.), and the use of VR is just one of them. Pedagogist Veronica Pantelidis expressed her opinion on the conditions that recommend VR, especially for learning and teaching. Based on the points reported in her analysis [60,61], some considerations founded by the authors during the final development phase are here reported. The development of an XR environment can be useful and used effectively:

- if the simulation as an alternative to the real environment allows for greater, more intuitive and faster learning,
- if the interaction with a model is more motivating than the interaction with reality,
- if you travel, costs or logistical difficulties in reaching the site make virtual reality more convenient,
- if the experience of creating a simulated environment or model is important to achieve learning objectives,
- if the visualisation of information and its manipulation using graphic symbols and the latest generation tools can be more easily understood, making the imperceptible perceptible,
- if it is necessary to develop a participatory environment that can only exist if generated with a computer,
- if it is necessary to give disabled people the opportunity to experiment, which they could not do otherwise,

The conditions that advise against the use of virtual reality in teaching are the following:

- whether the “real” learning environment is available and accessible,
- if interaction with real humans, professionals, tutors, teachers, students is necessary,
- if the use of a virtual environment can be physically or emotionally damaging,
- if the use of a virtual environment can provoke a simulation so convincing as to lead some participants to confuse the model with reality,
- if virtual reality is too expensive to justify in light of the expected results.

In addition to these pros and cons of a general nature and applicable to possible future developments in this field, the Arco della Pace case study has highlighted how the XR has become an innovative tool thanks to its multisensory and engaging nature, satisfying the principles of active learning. In fact, immersive virtual experiences have favoured the sense of presence and embodiment, both key factors capable of promoting learning and knowledge of intangible values such as the historical and cultural background represented in the decorative apparatus of the monument.

Learning these values became an active process in which the person builds his knowledge by extracting meanings from interactions with the surrounding virtual world. Thanks to interacting and extracting meanings from the objects surrounding him, the user creates mental models to understand reality. Consequently, the proposed virtual path became a dynamic process in which the person is the protagonist and active participant in the learning process. In turn, it has fostered an emotionally positive experience of involvement, promoting the onset and maintenance of high levels of attention and concentration.

On the other hand, it is also essential to consider the consequences of using virtual reality in health. In fact, some studies have found numerous problems related to what is called “cybersickness”, or symptoms of motion sickness due to diving. Participants sometimes reported experiencing headaches, nausea, disorientation and vision problems. Accordingly, the levels of interactivity developed had to deal with requirements that made them capable of not incurring the issues recently reported by the first manufacturers of gaming platforms. Consequently, the XR projects proposed in this study had to avoid VR sickness (dizziness, nausea, disorientation, sweating, and others). One of the main factors that can affect this is the framerate dropping too low. In Table 5, the recommended framerates for several of the VR headsets that Unreal Engine supports are reported:

**Table 5.** Specification to avoid VR sickness.

Device	Frame Per Second (FPS)
Vive	90
Gear VR	60
PSVR	Variable up to 120
Rift Retail	90
DK1	0
DK 2	75

Finally, VR environments require the utmost attention from the user (who is immersed in a reconstructed environment within which he can move and interact only “digitally”), making the technology inadequate for interaction. In this context, AR makes it possible to integrate the experience perfectly into the daily interactions that users have in the real world, facilitating collaboration between teams located in different places or accelerating digital learning, design and innovation processes. For these reasons, the development of a web-based AR library has made it possible to increase the usefulness of digital models, defining a new way of sharing information and objects simultaneously. On the other hand, the development of the library itself inevitably had to face specific requirements that led to a reduction in terms of LOD and LOI, such as

- the formats to be used (FBX, OBJ),
- the limited size of the shared models in terms of bytes (50,100,200 MB),
- the reduction of the result of the textures associated with the models (value to be considered in the general size of the AR object),
- compatibility with web browsers (desktop and mobile), and
- navigation and controls (Interface, Orbit Mode and First-Person Mode)

## 6. Discussion and Conclusions

An XR project has been developed to share the model and the related tangible and intangible values. Computer vision and imaging processing allow authors to improve the information mapping and sharing of the scan-to-HBIM process, creating novel XR environments containing the history of the monuments, high-resolution models of the statuary and the decorative apparatus and interactive virtual objects (IVO). New exchange formats, new game engine platforms, and visual scripting were used to complete the architectural study of such an important monument in the context of Milan. Right from the start, the work carried out aimed to certify how XR relates to the memory and custody of the built heritage, in particular the Arch of Peace in Milan, which in recent years has had less and less maintenance and restorations. Thanks to technological evolution, the virtual experience has become indispensable for the enjoyment of the architectural, artistic and cultural heritage to an increasingly large audience. In fact, this method shows how integration between the scan-to-BIM process, HBIM and XR allows users, in addition to painstakingly and faithfully recomposing any type of object, in this architectural case, to implement the knowledge of our built heritage. XR sees an infinite application in many fields, and its use in the context of cultural heritage and built heritage has an

enormous development prospect. Moreover, Italy contains an extensive cultural heritage within its territory, and the possible digitisation of this capital is one of the primary purposes of researchers, scholars, and experts in the sector. As previously mentioned, a fundamental role is recognised in continuous technological development, which contributes to improving virtual reality experiences, making the generative process and the subsequent governmental act even more immediate.

**Author Contributions:** Conceptualization, F.B. and A.M.; methodology, F.B. and A.M.; software, F.B. and A.M.; validation, F.B. and A.M.; F.B. and A.M.; investigation, F.B. and A.M.; resources, F.B. and A.M.; data curation, F.B. and A.M.; writing—original draft preparation, F.B. and A.M.; writing—review and editing, F.B. and A.M.; visualization, F.B. and A.M.; supervision, F.B. and A.M.; project administration, F.B. and A.M.; funding acquisition, F.B. and A.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Data will be available upon request.

**Acknowledgments:** The authors thank the Municipality of Milan (Italy), ENAC and Michela Sartori Prefettura of Milan Area I—Ordine e Sicurezza Pubblica for the authorizations to fly. The authors thank Jacopo Alberto Bonini and Mohamed Jaabar and Luca Melotto for the content implementation process of the virtual museum of Arco della Pace, Milan, Italy.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Rahaman, H.; Champion, E. To 3D or Not 3D: Choosing a Photogrammetry Workflow for Cultural Heritage Groups. *Heritage* **2019**, *2*, 112. [[CrossRef](#)]
- Marín-Buzón, C.; Pérez-Romero, A.; López-Castro, J.L.; Jerbania, I.B.; Manzano-Agugliaro, F. Photogrammetry as a New Scientific Tool in Archaeology: Worldwide Research Trends. *Sustainability* **2021**, *13*, 5319. [[CrossRef](#)]
- Honarmand, M.; Shahriari, H. Geological Mapping Using Drone-Based Photogrammetry: An Application for Exploration of Vein-Type Cu Mineralization. *Minerals* **2021**, *11*, 585. [[CrossRef](#)]
- Zanutta, A.; Lambertini, A.; Vittuari, L. UAV Photogrammetry and Ground Surveys as a Mapping Tool for Quickly Monitoring Shoreline and Beach Changes. *J. Mar. Sci. Eng.* **2020**, *8*, 52. [[CrossRef](#)]
- Luchowski, L.; Pojda, D.; Tomaka, A.A.; Skabek, K.; Kowalski, P. Multimodal Imagery in Forensic Incident Scene Documentation. *Sensors* **2021**, *21*, 1407. [[CrossRef](#)] [[PubMed](#)]
- Paoli, A.; Neri, P.; Razonale, A.V.; Tamburrino, F.; Barone, S. Sensor Architectures and Technologies for Upper Limb 3D Surface Reconstruction: A Review. *Sensors* **2020**, *20*, 6584. [[CrossRef](#)] [[PubMed](#)]
- Rupnik, E.; Jansa, J.; Pfeifer, N. Sinusoidal Wave Estimation Using Photogrammetry and Short Video Sequences. *Sensors* **2015**, *15*, 30784–30809. [[CrossRef](#)] [[PubMed](#)]
- Nikolakopoulos, K.; Kyriou, A.; Koukouvelas, I.; Zygouri, V.; Apostolopoulos, D. Combination of Aerial, Satellite, and UAV Photogrammetry for Mapping the Diachronic Coastline Evolution: The Case of Lefkada Island. *ISPRS Int. J. Geo Inf.* **2019**, *8*, 489. [[CrossRef](#)]
- Mancini, F.; Salvini, R. Applications of photogrammetry for environmental research (Editorial). *ISPRS Int. J. Geo Inf.* **2020**, *8*, 542. [[CrossRef](#)]
- Burdziakowski, P. Increasing the Geometrical and Interpretation Quality of Unmanned Aerial Vehicle Photogrammetry Products using Super-Resolution Algorithms. *Remote Sens.* **2020**, *12*, 810. [[CrossRef](#)]
- Rocha, G.; Mateus, L.; Fernández, J.; Ferreira, V. A Scan-to-BIM Methodology Applied to Heritage Buildings. *Heritage* **2020**, *3*, 4. [[CrossRef](#)]
- Wang, Q.; Guo, J.; Kim, M.-K. An Application Oriented Scan-to-BIM Framework. *Remote Sens.* **2019**, *11*, 365. [[CrossRef](#)]
- Yang, Y.; Xu, C.; Dong, F.; Wang, X. A New Multi-Scale Convolutional Model Based on Multiple Attention for Image Classification. *Appl. Sci.* **2019**, *10*, 101. [[CrossRef](#)]
- Butt, F.S.; Blunda, L.L.; Wagner, M.F.; Schäfer, J.; Medina-Bulo, I.; Gómez-Ullate, D. Fall Detection from Electrocardiogram (ECG) Signals and Classification by Deep Transfer Learning. *Information* **2021**, *12*, 63. [[CrossRef](#)]
- Gochoo, M.; Rizwan, S.A.; Ghadi, Y.Y.; Jalal, A.; Kim, K. A Systematic Deep Learning Based Overhead Tracking and Counting System Using RGB-D Remote Cameras. *Appl. Sci.* **2021**, *11*, 5503. [[CrossRef](#)]
- Neptune, N.; Mothe, J. Automatic Annotation of Change Detection Images. *Sensors* **2021**, *21*, 1110. [[CrossRef](#)] [[PubMed](#)]
- Dore, C. Integration of HBIM and 3D GIS for Digital Heritage Modelling. In Proceedings of the Digital Documentation International Conference, Edinburgh, UK, 22–23 October 2012. [[CrossRef](#)]

18. Brumana, R.; Georgopoulos, A.; Oreni, D.; Raimondi, A.; Bregianni, A. HBIM for Documentation, Dissemination and Management of Built Heritage. The Case Study of St. Maria in Scaria d’Intelvi. *Int. J. Herit. Digit. Era* **2013**, *2*, 433–451. [[CrossRef](#)]
19. Inzerillo, L.; Lo Turco, M.; Parrinello, S.; Santagati, C.; Valenti, G.M.; Inzerillo, L. BIM and architectural heritage: Towards an operational methodology for the knowledge and the management of Cultural Heritage. *Disegnarecon* **2016**, *9*.
20. Chiabrando, F.; Sammartano, G.; Spanò, A.T. Historical buildings models and their handling via 3D survey: From points clouds to user-oriented HBIM. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *XLI-B5*, 633–640. [[CrossRef](#)]
21. Costantino, D.; Pepe, M.; Restuccia, A.G. Scan-to-HBIM for conservation and preservation of Cultural Heritage building: The case study of San Nicola in Montedoro church (Italy). *Appl. Geomat.* **2021**. [[CrossRef](#)]
22. Gironacci, I.M. State of the Art of Extended Reality Tools and Applications in Business. In *Transdisciplinary Perspectives on Risk Management and Cyber Intelligence*; IGI Global: Hershey, PA, USA, 2020; pp. 105–118.
23. Banfi, F. BIM orientation: Grades of generation and information for different type of analysis and management process. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 57–64. [[CrossRef](#)]
24. Roberts, C.J.; Pärn, E.A.; Edwards, D.J.; Aigbavboa, C. Digitalising asset management: Concomitant benefits and persistent challenges. *Int. J. Build. Pathol. Adapt.* **2018**, *36*, 152–173. [[CrossRef](#)]
25. Visintini, D.; Marcon, E.; Pantò, G.; Canevese, E.P.; De Gottardo, T.; Bertani, I. Advanced 3d modeling versus building information modeling: The case study of palazzo ettoreo in sacile (Italy). *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *42*, 1137–1143. [[CrossRef](#)]
26. Brumana, R.; Ioannides, M.; Previtali, M. Holistic heritage building information modelling (hhbim): From nodes to hub networking, vocabularies and repositories. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *42*, 309–316. [[CrossRef](#)]
27. López, F.J.; Lerones, P.M.; Llamas, J.M.; Gómez-García-Bermejo, J.; Zalama, E. Linking HBIM graphical and semantic information through the Getty AAT: Practical application to the Castle of Torrelobatón. In Proceedings of the IOP Conference Series: Materials Science and Engineering; Institute of Physics Publishing: Bristol, UK, 2018; Volume 364.
28. Pauwels, P.; De Meyer, R.; Van Campenhout, J. Interoperability for the design and construction industry through semantic web technology. In Proceedings of the Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Cham, Switzerland, 2011; Volume 6725 LNCS, pp. 143–158.
29. Costa, G.; Madrazo, L. Connecting building component catalogues with BIM models using semantic technologies: An application for precast concrete components. *Autom. Constr.* **2015**, *57*, 239–248. [[CrossRef](#)]
30. Kang, T.W.; Choi, H.S. BIM perspective definition metadata for interworking facility management data. *Adv. Eng. Inform.* **2015**, *29*, 958–970. [[CrossRef](#)]
31. Niknam, M.; Karshenas, S. A shared ontology approach to semantic representation of BIM data. *Autom. Constr.* **2017**, *80*, 22–36. [[CrossRef](#)]
32. A Semantic Web Primer for Object-Oriented Software Developers. Available online: <https://www.w3.org/TR/sw-oosd-primer/> (accessed on 2 July 2021).
33. Pauwels, P.; Terkaj, W. EXPRESS to OWL for construction industry: Towards a recommendable and usable ifcOWL ontology. *Autom. Constr.* **2016**, *63*, 100–133. [[CrossRef](#)]
34. Simeone, D.; Cursi, S.; Toldo, I.; Carrara, G. B(H)IM -Built Heritage Information Modelling. In Proceedings of the 32nd eCAADe Conference, Newcastle Upon Tyne, UK, 10–12 September 2014; Volume 1, pp. 613–622.
35. Bruno, N.; Roncella, R. A restoration oriented HBIM system for cultural heritage documentation: The case study of Parma cathedral. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *42*, 171–178. [[CrossRef](#)]
36. Chiabrando, F.; Lo Turco, M.; Rinaudo, F. Modeling the decay in an HBIM starting from 3D point clouds. A followed approach for cultural heritage knowledge. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 605–612. [[CrossRef](#)]
37. Besl, P.J.; McKay, N.D. Method for registration of 3-D shapes. In Proceedings of the Sensor Fusion IV: Control Paradigms and Data Structures; SPIE: Boston, MA, USA, 1992; Volume 1611, pp. 586–606.
38. Altun, M.; Akcamete, A. A Method for facilitating 4D modeling by automating task information generation and mapping. In *Advances in Informatics and Computing in Civil and Construction Engineering*; Springer International Publishing: New York, NY, USA, 2019; pp. 479–486.
39. Cogima, C.; Paiva, P.; Dezen-Kempter, E.; Carvalho, M.A.G.; Soibelman, L. The role of knowledge-based information on BIM for Built Heritage. In *Advances in Informatics and Computing in Civil and Construction Engineering*; Springer International Publishing: New York, NY, USA, 2019; pp. 27–34.
40. Diara, F.; Rinaudo, F. From reality to parametric models of cultural heritage assets for HBIM. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *42*, 413–419. [[CrossRef](#)]
41. Fai, S.; Sydor, M. Building Information Modelling and the documentation of architectural heritage: Between the “typical” and the “specific”. In Proceedings of the Digital Heritage International Congress (Digital Heritage), Marseille, France, 28 October–1 November 2013; Volume 1, pp. 731–734.
42. Heesom, D.; Boden, P.; Hatfield, A.; Roobie, S.; Andrews, K.; Berwari, H. Developing a collaborative HBIM to integrate tangible and intangible cultural heritage. *Int. J. Build. Pathol. Adapt.* **2021**, *39*, 72–95. [[CrossRef](#)]
43. Ioannides, M.; Magnenat-Thalmann, N.; Papagiannakis, G. *Mixed Reality and Gamification for Cultural Heritage*; Springer International Publishing: New York, NY, USA, 2017.

44. Quattrini, R.; Pierdicca, R.; Morbidoni, C. Knowledge-based data enrichment for HBIM: Exploring high-quality models using the semantic-web. *J. Cult. Herit.* **2017**, *28*, 129–139. [[CrossRef](#)]
45. Oreni, D.; Brumana, R.; Georgopoulos, A.; Cuca, B. Hbim for conservation and management of built heritage: Towards a library of vaults and wooden beam floors. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2013**, *2*, 215–221. [[CrossRef](#)]
46. Bagnolo, V.; Argiolas, R. Scan-to-BIM process versus 3D procedural modelling of gothic masonry vaults. In *Springer Tracts in Civil Engineering*; Springer: New York, NY, USA, 2021; pp. 17–31.
47. Jang, J.; Ko, Y.; Shin, W.S.; Han, I. Augmented Reality and Virtual Reality for Learning: An Examination Using an Extended Technology Acceptance Model. *IEEE Access* **2021**, *9*, 6798–6809. [[CrossRef](#)]
48. Jung, K.; Nguyen, V.T.; Lee, J. Blocklyxr: An interactive extended reality toolkit for digital storytelling. *Appl. Sci.* **2021**, *11*, 1073. [[CrossRef](#)]
49. Lerma, J.L.; Navarro, S.; Cabrelles, M.; Villaverde, V. Terrestrial laser scanning and close range photogrammetry for 3D archaeological documentation: The Upper Palaeolithic Cave of Parpalló as a case study. *J. Archaeol. Sci.* **2010**, *37*, 499–507. [[CrossRef](#)]
50. Loaiza Carvajal, D.A.; Morita, M.M.; Bilmes, G.M. Virtual museums. Captured reality and 3D modeling. *J. Cult. Herit.* **2020**, *45*, 234–239. [[CrossRef](#)]
51. Pybus, C.; Graham, K.; Doherty, J.; Arellano, N.; Fai, S. New realities for Canada’s parliament: A workflow for preparing heritage BIM for game engines and virtual reality. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *42*, 945–952. [[CrossRef](#)]
52. Trunfio, M.; Lucia, M.D.; Campana, S.; Magnelli, A. Innovating the cultural heritage museum service model through virtual reality and augmented reality: The effects on the overall visitor experience and satisfaction. *J. Herit. Tour.* **2020**. [[CrossRef](#)]
53. Stöcker, C.; Eltner, A.; Karrasch, P. Measuring gullies by synergetic application of UAV and close range photogrammetry—A case study from Andalusia, Spain. *Catena* **2015**, *132*, 1–11. [[CrossRef](#)]
54. Ente Nazionale per l’Aviazione Civile—Italian Civil Aviation Authority. Available online: <https://www.enac.gov.it/> (accessed on 2 July 2021).
55. d-flight—Enabling Autonomous Flight. Available online: [https://www.d-flight.it/new\\_portal/](https://www.d-flight.it/new_portal/) (accessed on 2 July 2021).
56. Banfi, F. HBIM generation: Extending geometric primitives and BIM modelling tools for heritage structures and complex vaulted systems. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *42*, 139–148. [[CrossRef](#)]
57. Brumana, R.; Stanga, C.; Banfi, F. Models and scales for quality control: Toward the definition of specifications (GOA-LOG) for the generation and re-use of HBIM object libraries in a Common Data Environment. *Appl. Geomat.* **2021**. [[CrossRef](#)]
58. Lo Turco, M.; Calvano, M.; Giovannini, E.C. Data modeling for museum collections. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *42*, 433–440. [[CrossRef](#)]
59. Javidi, G. *Virtual Reality and Education*; University of South Florida: Tampa, FL, USA, 1999.
60. Pantelidis, V. Virtual Reality in the Classroom. *Educ. Technol.* **1993**, *33*, 23–27.
61. Pantelidis, V.S. Reasons to Use Virtual Reality in Education and Training Courses and a Model to Determine When to Use Virtual Reality. *Themes Sci. Technol. Educ.* **2010**, *2*, 59–70.





Article

# Comparing CAM Algorithms for the Identification of Salient Image Features in Iconography Artwork Analysis

Nicolò Oreste Pincioli Vago <sup>1,2</sup>, Federico Milani <sup>1,\*</sup>, Piero Fraternali <sup>1</sup> and Ricardo da Silva Torres <sup>2</sup>

<sup>1</sup> Department of Electronics Information and Bioengineering, Politecnico di Milano, 20133 Milano, Italy; nicoloreste.pincioli@mail.polimi.it (N.O.P.V.); piero.fraternali@polimi.it (P.F.)

<sup>2</sup> Department of ICT and Engineering, NTNU—Norwegian University of Science and Technology, 6009 Ålesund, Norway; nicoloop@stud.ntnu.no (N.O.P.V.); ricardo.torres@ntnu.no (R.d.S.T.)

\* Correspondence: federico.milani@polimi.it

**Abstract:** Iconography studies the visual content of artworks by considering the themes portrayed in them and their representation. Computer Vision has been used to identify iconographic subjects in paintings and Convolutional Neural Networks enabled the effective classification of characters in Christian art paintings. However, it still has to be demonstrated if the classification results obtained by CNNs rely on the same iconographic properties that human experts exploit when studying iconography and if the architecture of a classifier trained on whole artwork images can be exploited to support the much harder task of object detection. A suitable approach for exposing the process of classification by neural models relies on Class Activation Maps, which emphasize the areas of an image contributing the most to the classification. This work compares state-of-the-art algorithms (CAM, Grad-CAM, Grad-CAM++, and Smooth Grad-CAM++) in terms of their capacity of identifying the iconographic attributes that determine the classification of characters in Christian art paintings. Quantitative and qualitative analyses show that Grad-CAM, Grad-CAM++, and Smooth Grad-CAM++ have similar performances while CAM has lower efficacy. Smooth Grad-CAM++ isolates multiple disconnected image regions that identify small iconographic symbols well. Grad-CAM produces wider and more contiguous areas that cover large iconographic symbols better. The salient image areas computed by the CAM algorithms have been used to estimate object-level bounding boxes and a quantitative analysis shows that the boxes estimated with Grad-CAM reach 55% average IoU, 61% GT-known localization and 31% mAP. The obtained results are a step towards the computer-aided study of the variations of iconographic elements positioning and mutual relations in artworks and open the way to the automatic creation of bounding boxes for training detectors of iconographic symbols in Christian art images.



**Citation:** Pincioli Vago, N.O.; Milani, F.; Fraternali, P.; da Silva Torres, R. Comparing CAM Algorithms for the Identification of Salient Image Features in Iconography Artwork Analysis. *J. Imaging* **2021**, *7*, 106. <https://doi.org/10.3390/jimaging7070106>

Academic Editors: Giovanna Castellano, Gennaro Vessio and Fabio Bellavia

Received: 15 May 2021

Accepted: 24 June 2021

Published: 29 June 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Iconography is the discipline that concerns itself with the subject matter of artworks, as opposed to their form [1]. It is studied to understand the meaning of artworks and to analyze the influence of culture and beliefs on art representations across the world, from the Nasca [2] to the Byzantine [3] civilization. Iconography is a prominent topic of the art history studied through centuries [4–6]. The attribution of iconographic elements (henceforth *classes*) is an important task in art history, related to the interpretation of meaning and to the definition of the geographical and temporal context of an artwork.

With the advent of digital art collections, iconographic class attribution has acquired further importance, as a way to provide a significant index on top of digital repositories of art images, supporting both students and experts in finding and comparing works by their iconographic attributes. However, the analysis of iconography requires specialized skills, based on the deep knowledge of the symbolic meaning of a very high number of

elements and of their evolution in space and time. The WikiPedia page on Christian Saint symbolism ([https://en.wikipedia.org/wiki/Saint\\_symbolism](https://en.wikipedia.org/wiki/Saint_symbolism)—accessed on 15 May 2021) lists 257 characters with 791 attributes. This makes the manual attribution of iconographic classes to image collections challenging, due to the tension between the available amount of expert work and the high number of items to be annotated.

A viable alternative relies on the use of semi-automatic computer-aided solutions supporting the expert annotator in the task of associating iconographic classes to art images. Computer Vision (CV) has already been used for artwork analysis tasks, such as genre identification [7], author identification [8], and even subject identification and localization [9]. The field of computer-aided iconographic analysis is more recent and addressed by few works [10,11]. Borrowing the standard CV terminology, the problem of computer-aided iconographic analysis can be further specialized into iconography classification, which tackles the association of iconographic classes to an artwork image as a whole, and iconography detection, which addresses the identification of the regions of an image in which the attributes representing an iconographic class appear.

Applying CV to the analysis of art iconographic poses challenges, in part, general and, in part, specific to the art iconography field. As in general-purpose image classification and object detection, the availability of large high quality training data is essential. The natural image dataset in use today are very large and provided with huge numbers of annotations. Conversely, in the narrower art domain, image datasets are less abundant, smaller, and with less high-quality annotations. Furthermore, unlike natural images, painting images are characterized by less discriminative features than natural ones. The color palette is more restricted and subject to artificial effects, such as colored shadows and chiaroscuro. Images of paintings may also portray partially deteriorated subjects (e.g., in frescoes) and belong to historical archives of black and white photos.

Despite the encouraging results of applying Convolutional Neural Networks (CNNs) for iconography classification [11], it remains unclear how such a task is performed by artificial models. Depending on the class, the human expert may consider the whole scene portrayed in the painting or instead focus on specific hints. Considering Christian art iconography, an example of the first scenario occurs in paintings of complex scenes such as the crucifixion or the visitation of the magi. The latter case is typical of the identification of characters, especially Christian saints, which depends on the presence of very distinctive attributes. When CNNs are used for the classification task, the problem of explainability arises, i.e., of exposing how the CNN has produced a given result. A widely used strategy to clarify CNN image classification results relies on the use of Class Activation Maps [12–14], which visualize the regions of the input images that have the most impact on the prediction of the CNN. Computing the most salient regions of an image with respect to its iconography can help automate the creation of bounding boxes around the significant elements of an artwork from image-wide annotations only. This result could reduce the effort of building training sets for the much harder task of iconography detection.

This paper addresses the following research questions:

- Are CAMs an effective tool for understanding how a CNN classifier recognizes the iconographic classes of a painting?
- Are there significant differences in the state-of-the-art CAM algorithms with respect to their ability to support the explanation of iconography classification by CNNs?
- Are the image areas highlighted by CAMs a good starting point for creating semi-automatically the bounding boxes necessary for training iconography detectors?

The contributions of the paper can be summarized as follows:

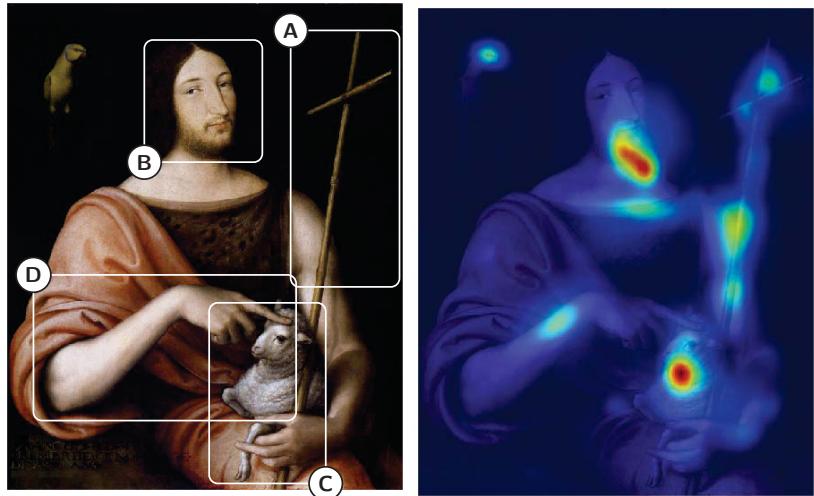
- We apply four state-of-the-art class activation map algorithms (namely, CAM [15], Grad-CAM [16], Grad-CAM++ [17], and Smooth Grad-CAM++ [18]) to the CNN iconography classification model presented in [11], which exploits a backbone based on ResNet50 [19] trained on the ImageNet dataset [20] and refined on the ArtDL dataset (<http://www.artdl.org>—accessed on 15 May 2021) consisting of 42,479 images of artworks portraying Christian Saints divided into 10 classes. Note that, in order

to avoid ambiguity, we refer to the specific algorithm as “CAM” and to the generic output as “class activation maps”.

- For the quantitative evaluation of the different algorithms, a test dataset has been built which comprises 823 images annotated with 2957 bounding boxes surrounding specific iconographic symbols. One such annotated image is shown in Figure 1. We use the Intersection over Union (IoU) metrics to measure the agreement between the areas of the image highlighted by the algorithm and those annotated manually as ground truth. Furthermore, we analyze the class activation map area based on percentage of covered bounding boxes and percentage of covered area that does not contain any iconographic symbol.
- The comparison shows that Grad-CAM, Grad-CAM++, and Smooth Grad-CAM++ deliver better results than the original CAM algorithm in terms of area coverage and explainability. This finding confirms the result discussed in [18] for natural images. Smooth Grad-CAM++ produces multiple disconnected image regions that identify small iconographic symbols quite precisely. Grad-CAM produces wider and more contiguous areas that cover well both large and small iconographic symbols. To the best of our knowledge, such a comparison has not been performed before in the context of artwork analysis.
- We perform a qualitative evaluation by examining the overlap between the ground-truth bounding boxes and the class activation maps. This investigation illustrates the strengths and weaknesses of the analyzed algorithms, highlights their capacity of detecting symbols that were missed by the human annotator and discusses cases of confusion between the symbols of different classes. A simple procedure is tested for selecting “good enough” class activation maps and for creating symbol bounding boxes automatically from them. The results of such a procedure are illustrated visually.
- We deepen the evaluation by measuring quantitatively the agreement between the ground-truth bounding boxes and the bounding boxes estimated from the class activation maps. The assessment shows that the whole Saint bounding boxes computed from the Grad-CAM class activation maps obtain 55% average IoU, 61% GT-known localization and 31% mAP. Such results obtained by a simple post-processing of the output of a general purpose CNN interpretability technique pave the way to the use of automatically computed bounding boxes for training weakly supervised object detectors in artwork images.

Figure 1 shows an example of the assessment performed in this paper. On the left, an image of Saint John the Baptist has been manually annotated with the regions (from A to D) associated with key symbols relevant for iconography classification. On the right, the same image is overlaid with the CAM heat map showing the regions contributing the most to the classification.

The rest of the paper is organized as follows: Section 2 surveys related work; Section 3 describes the different CAM variants considered in our study; Section 4 describes the adopted evaluation protocol and the results of the quantitative and the qualitative analysis; finally, Section 5 draws the conclusions and outlines possible future work.



**Figure 1.** On the left: Saint John the Baptist image and iconographic symbols identified manually (e.g., cross (A), face (B), and lamb (C), and hand pointing at lamb (D)). On the right: the CAM heat map associated with classification results of a CNN-based solution.

## 2. Related Work

This section surveys the essential previous research in the fields of automated artwork analysis and CNN interpretability that are the foundations of our work.

### 2.1. Automated Artwork Image Analysis

The large availability of artworks in digital format has allowed researchers to perform automated analysis in the fields of digital humanities and cultural heritage by means of Computer Vision and Deep Learning methods. Several datasets containing various types of artworks have been proposed to support such studies [10,11,21–26].

The performed analyses span several classification tasks and techniques: from style classification to artist identification, comprising also medium, school, and year classification [27–29]. These researches are useful to support cultural heritage studies and asset management, e.g., automatic cataloguing of unlabeled works in online and museum collections, but their results can be exploited for more complex applications, such as authentication, stylometry [30], and forgery detection [31].

A task that is more related to our proposal is artwork content analysis, which focuses on the automatic identification and, if possible, localization of objects inside artworks. The literature contains several state-of-the-art approaches [9–11,32–35]. Since there is abundance of deep learning models trained with natural images but a deficiency of art-specific models, many studies focus on the transferability of previous knowledge to the art domain [11,35–38]. This approach is known as Transfer Learning and consists in fine-tuning a network, previously trained with natural images, using art images. The consensus is that Transfer Learning is beneficial for tasks related to artworks analysis.

### 2.2. Interpretability and Activation Maps

In recent years, Deep Learning models have been treated as black-boxes, i.e., architectures that do not expose their internal operations to the user. These systems are used for various approaches and their interpretability is fundamental in many fields, especially when the outputs of the models are used for sensitive applications. In the literature, there are many techniques that aim at explaining the behavior of neural models [39,40]. Saliency Masks are used to address the outcome explanation problem by providing a visualization of which part of the input data is mainly responsible for the network prediction. The most

popular Saliency Masks are obtained with the Class Activation Map (CAM) approach. CAMs [15] have shown their effectiveness in highlighting the most discriminative areas of an image in several fields, ranging from medicine [41] to fault diagnostics [12]. The original formulation of CAMs has been subsequently improved. Selvaraju et al. [16] introduced Grad-CAM, which exploits the gradients that pass through the final convolutional layer to compute the most salient areas of the input. Chattopadhyay et al. [17] introduced Grad-CAM++ which considers gradients too but is based on a different mathematical formulation that improves the localization of single and multiple instances. Smooth Grad-CAM++ [18] applies Grad-CAM++ iteratively on the combination of the original image and a Gaussian noise.

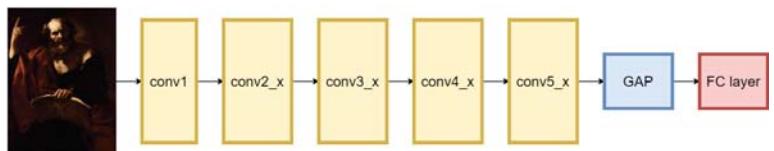
The use of CAMs is not limited to the explainability of Deep Learning classification models but is the starting point for studies related to the weakly supervised localization of content inside the images [42].

This paper focuses on the comparison of different CAM algorithms on the task of iconography classification to determine which variant may be more suitable for weakly supervised studies. Since CAM algorithms are most often studied only for natural images, the aim of the work is also to address the research gap about the utility of CAMs for the art domain.

### 3. Class Activation Maps for Iconography Classification

This paper compares different CAM algorithms: Grad-CAM, Grad-CAM++, and Smooth Grad-CAM++. Their implementation is based on the mathematical definitions provided, respectively, by [15–18].

Figure 2 shows the ResNet50 classifier architecture used to compute the class activation maps. The input of the network is an image and the output is the set of probabilities associated with the different classes. In the evaluation, the input images portray art works and the output classes denote 10 Christian Saints. ResNet50 contains an initial convolutional layer (conv1) followed by a sequence of convolutional residual blocks (conv2\_x … conv5\_x). A Global Average Pooling (GAP) module computes the average value for each feature map obtained as an output of the last layer (conv5\_x). The probability estimates are computed by the last component, which is typically a fully connected (FC) layer [43].



**Figure 2.** The ResNet50 architecture.

#### 3.1. CAM

CAMs [15] are based on the use of GAP, which has been demonstrated to have remarkable localization abilities [44]. The GAP operation averages the feature maps of the last convolutional layer and feeds the obtained values to the final fully connected layer that performs the actual classification. Class activation maps are generated by performing a weighted sum of the feature maps of the last convolutional layer for each class. The actual class activation map value  $M_c(x, y)$  for a class  $c$  and a position  $x, y$  in the input image is expressed as follows:

$$M_c(x, y) = \sum_k w_k^c A_k(x, y) \quad (1)$$

where  $A_k(x, y)$  is the activation value of feature map  $k$  in the last convolutional layer at position  $(x, y)$ , and  $w_k^c$  is the weight associated with feature map  $k$  and with class  $c$ . Intuitively, a high CAM value at position  $x, y$  is the result of an average high activation value of all the feature maps of the last convolutional layer.

Differently from the original approach, we compute the CAM output not only for the predominant class, but for all the classes. The ArtDL dataset contains multi-class multi-label images and this formulation allows us to analyze which regions of the artwork are associated with which classes, also in the case of wrong classification.

### 3.2. Grad-CAM

Grad-CAM [16] is a variant of CAM which considers not only the weights but also the gradients flowing into the last convolution layer. In this way, also the layers preceding the last one contribute to the activation map. An advantage of using gradients is that Grad-CAM can be applied to any layer of the network. Still, the last one is especially relevant for the localization of the parts of the image that contribute most to the final prediction. Furthermore, the layer used as input for the prediction can be followed by any module and not only by a fully connected layer. Grad-CAM exploits the parameters  $\alpha_c^k$ , which represents the neuron importance weights and are calculated as:

$$\alpha_c^k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (2)$$

where  $\frac{1}{Z} \sum_i \sum_j$  denotes the global average pooling operation ( $Z = i \cdot j$ ) and  $\frac{\partial y^c}{\partial A_{ij}^k}$  denotes the back-propagation gradients. In the gradient expression,  $y^c$  is the score of the class  $c$  and  $A^k$  represents the  $k$ -th feature map. The Grad-CAM for a class  $c$  at position  $(x, y)$  is then given by:

$$M_{Grad-CAM}^c(x, y) = ReLU \left( \sum_k \alpha_c^k A^k(x, y) \right) \quad (3)$$

where the ReLU operator maps the negative values to zero. As in the case of CAM, we compute the output of Grad-CAM for all the classes under analysis.

### 3.3. Grad-CAM++

Grad-CAM++ [17] is a generalization of Grad-CAM aimed at better localizing multiple class instances and at capturing objects more completely. Differently from Grad-CAM, Grad-CAM++ applies a weighted average of the partial derivatives, with the purpose of covering a wider portion of the object. Given a class  $c$  with a score  $Y^c$  and the activation map  $A_{ij}^k$  calculated in the last convolutional layer, a parameter  $\alpha_{ij}^{kc}$  can be defined as follows:

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2}}{2 \frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2} + \sum_a \sum_b A_{ab}^k \left\{ \frac{\partial^3 Y^c}{(\partial A_{ij}^k)^3} \right\}} \quad (4)$$

The parameter  $w_k^c$ , which has the same role of  $\alpha_k^c$  in Grad-CAM, is defined as:

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} ReLU \left( \frac{\partial Y^c}{\partial A_{ij}^k} \right) \quad (5)$$

which leads to

$$w_k^c = \sum_{i,j} \left[ \frac{\frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2}}{2 \frac{\partial^2 Y^c}{(\partial A_{ij}^k)^2} + \sum_{a,b} A_{ab}^k \left\{ \frac{\partial^3 Y^c}{(\partial A_{ij}^k)^3} \right\}} \right] ReLU \left( \frac{\partial Y^c}{\partial A_{ij}^k} \right) \quad (6)$$

As in the other CAMs, it holds that

$$M_{Grad-CAM++}^c(x, y) = ReLU \left( \sum_k w_k^c A^k(x, y) \right) \quad (7)$$

### 3.4. Smooth Grad-CAM++

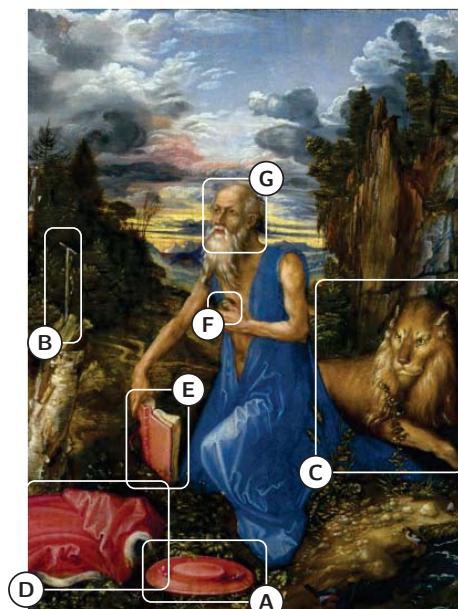
Smooth Grad-CAM++ [18] is a variant of Grad-CAM++ that can focus on subsets of feature maps or of neurons for identifying anomalous activations. Smooth Grad-CAM++ applies random Gaussian perturbations on the image  $z$  and exploits the visual sharpening of the class activation maps by averaging random samples taken from a feature map close to the input. The value of the activation map  $M_c$  in a position  $(x, y)$  is defined as:

$$M_{c(x,y)}(z) = \frac{1}{n} \sum_1^n M_{c(x,y)}^{GCpp}(z + \mathcal{N}(0, \sigma^2)) \quad (8)$$

where  $n$  is the number of samples,  $\mathcal{N}(0, \sigma^2)$  is the 0-mean Gaussian noise with standard deviation  $\sigma$ , and  $M_c^{GCpp}$  is the activation map for the input  $z + \mathcal{N}(0, \sigma^2)$ . The final result is obtained by iterating the computation of Grad-CAM++ on inputs resulting from the overlap of the original image and a random Gaussian noise.

### 4. Evaluation

The evaluation exploits the ArtDL dataset [11], an existing artwork collection annotated with image-level labels. The purpose of the evaluation is: (1) to understand whether the class activation maps are effective in localizing both the whole representation of an iconographic class and the distinct symbols that characterize it (the attributes associated with the classes present in the ArtDL dataset are illustrated in [45] and listed in [46]); (2) to compare CAMs algorithms in their ability to do so. To evaluate the localization ability of class activation maps, a subset of the images have been annotated with bounding boxes framing iconographic symbols associated with each Saint. Figure 3 illustrates the symbols in a painting of Saint Jerome. The bounding boxes are used for the quantitative assessment of class activation maps algorithms with the metrics described in Section 4.5. A qualitative analysis is reported in Section 4.6.



**Figure 3. Saint Jerome**—The cardinal’s galero (A), the crucifix (B), the lion (C), the cardinal’s vest (D), the book (E), the stone in the hand (H), and the face (G).

#### 4.1. Dataset

The ArtDL dataset [11] comprises images of paintings that represent the Iconclass [47] categories of 10 Christian Saints: Saint Dominic, Saint Francis of Assisi, Saint Jerome, Saint John the Baptist, Saint Anthony of Padua, Saint Mary Magdalene, Saint Paul, Saint Peter, Saint Sebastian, and the Virgin Mary. The representation of such classes in Christian art paintings exploit specific symbols, i.e., markers that hint at the identity of the portrayed character. Table 1 summarizes the symbols associated with the 10 Iconclass categories represented in the ArtDL dataset.

**Table 1.** Iconclass categories and symbols associated with them.

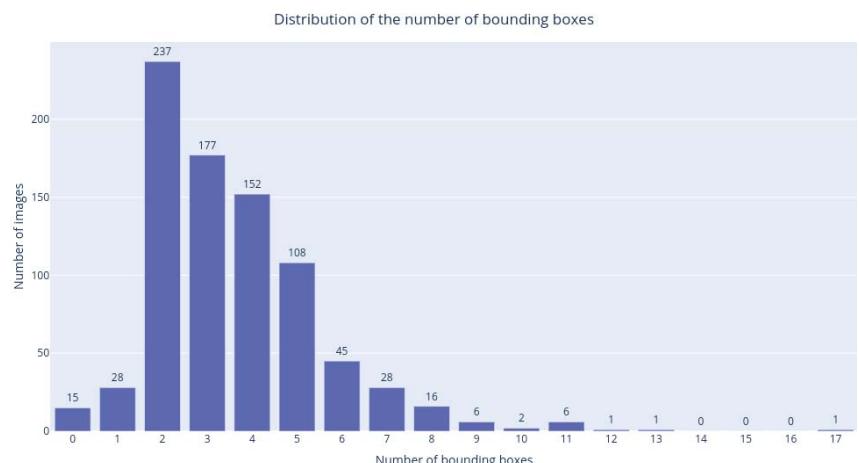
Iconclass Category	Symbols
Anthony of Padua	Baby Jesus, bread, book, lily, face, cloth
Dominic	Rosary, star, dog with a torch, face, cloth
Francis of Assisi	Franciscan cloth, wolf, birds, fish, skull, stigmata, face, cloth
Jerome	Hermitage, lion, cardinal's galero, cardinal vest, cross, skull, book, writing material, stone in hand, face, cloth
John the Baptist	Lamb, head on platter, animal skin, pointing at Christ, pointing at lamb, cross, face, cloth
Mary Magdalene	Ointment jar, long hair, washing Christ's feet, skull, crucifix, red egg, face, cloth
Paul	Sword, book, scroll, horse, beard, balding head, face, cloth
Peter	Keys, boat, fish, rooster, pallium, papal vest, inverted cross, book, scroll, bushy beard, bushy hair, face, cloth
Sebastian	Arrows, crown, face, cloth
Virgin Mary	Baby Jesus, rose, lily, heart, seven swords, crown of stars, serpent, rosary, blue robe, sun and moon, face, cloth, crown

The ArtDL images are associated with high-level annotations specifying which Iconclass categories appear in them (from a minimum of 1 to a maximum of 7). Whole-image labels are not sufficient to assess the different ways in which the class activation maps methods focus on the image content. For this purpose, it is necessary to annotate the dataset with bounding boxes that localize the symbols listed in Table 1. Out of the whole dataset, 823 sample images were selected and manually annotated with bounding boxes that frame each symbol separately. A symbol can either be included completely within a single bounding box (e.g., Saint Jerome's lion) or be split into multiple bounding boxes (e.g., Saint Peter's bushy hair, which are usually divided in two parts separated by the forehead). We consider a symbol representation as the union of all the bounding boxes annotated with the same symbol label. For instance, Saint Sebastian's arrows correspond to a unique symbol but are annotated with multiple bounding boxes. When the same symbol relates to multiple saints (e.g., Baby Jesus may appear with both the Virgin Mary and St. Anthony of Padua), its presence is denoted with a label composed of the symbol name and the Saint's name. While some symbols appear in the majority of the images of the corresponding Saint, others are absent or rarely present. For each Saint, only the symbols that appear in at least 5% of the paintings depicting the respective Saint are kept. This filter eliminates 23 of the 84 possible symbols associated with the 10 Iconclass categories and reduces the number of symbol bounding boxes from 2957 to 2887. Table 2 summarizes the characteristics of the dataset used to compare the class activation maps algorithms.

Figure 4 shows the distribution of the bounding boxes within the images. Most images contain from 2 to 5 bounding boxes and a few images do not contain any annotation. The latter case occurs when the automatic classification of the ArtDL dataset is incorrect (e.g., for images in which a character named Mary was incorrectly associated with the Virgin Mary).

**Table 2.** Symbol and bounding box distribution.

Iconclass Category	Symbol Classes	Symbol Bounding Boxes
<b>Anthony of Padua</b>	6	83
<b>Dominic</b>	4	59
<b>Francis of Assisi</b>	5	295
<b>Jerome</b>	11	434
<b>John the Baptist</b>	5	231
<b>Mary Magdalene</b>	5	283
<b>Paul</b>	6	132
<b>Peter</b>	9	408
<b>Sebastian</b>	3	267
<b>Virgin Mary</b>	7	695

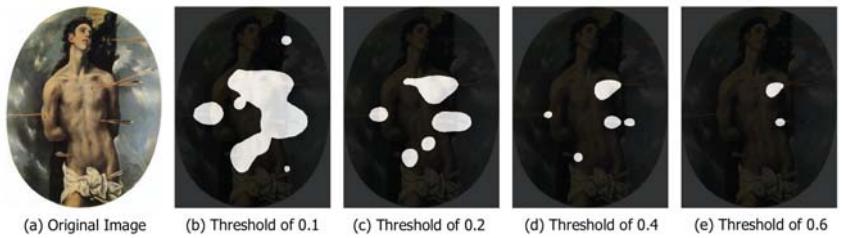
**Figure 4.** Bounding box distribution: most images contain from 2 to 5 bounding boxes (average = 3).

#### 4.2. Class Activation Maps Generation

The class activation maps are generated by feeding the image to the ResNet50 model and applying the computation explained in Section 3. They have a size equal to  $h \times w \times c$  where  $h$  and  $w$  are the height and width of the  $\text{conv5\_}x$  layer and  $c$  is the number of classes. Since the output size ( $h, w$ ) is smaller than the input size, due to the convolution operations performed by the ResNet architecture, each class activation map is upsampled with bilinear interpolation to match the input image size. Min-max scaling is applied to the upsampled class activation maps to normalize them in the  $[0, 1]$  range.

#### 4.3. Choice of the Threshold Value

A class activation map contains values in the range from 0 to 1. Given a threshold  $t$ , it is possible to separate the class activation map into background (pixels with a value lower than  $t$ ) and foreground (pixels with a value greater than  $t$ ). The choice of the threshold value aims at making foreground areas concentrate on the Saints' figure and symbols. Figure 5 shows the impact of applying different threshold values to a class activation map. As the threshold value increases, the foreground areas (in white) become smaller and more distinct and the background pixels increase substantially at the cost of fragmenting the foreground areas and missing relevant symbols. To investigate the choice of the proper threshold, the quantitative evaluation of Section 4.5 reports results obtained with multiple values uniformly distributed from 0 to 1 with a step of 0.05.



**Figure 5.** Analysis with different thresholds—black areas correspond to class activation map values below the specified threshold (background) while white pixels correspond to class activation map values greater or equal than the threshold (foreground). An increment in the threshold value results in smaller and more distinct areas. Original image (a), cam with threshold at 0.1 (b), cam with threshold at 0.2 (c), cam with threshold at 0.4 (d), cam with threshold at 0.6 (e).

#### 4.4. Intersection Over Union Metrics

Intersection Over Union (IoU) is a standard metric used to compute the overlap between two different areas. It is defined as:

$$IoU = \frac{A_{\cap}}{A_{\cup}},$$

where  $A_{\cap}$  is the intersection between the two areas and  $A_{\cup}$  is their union. IoU ranges between 0 and 1, with 0 meaning that the two areas are disjoint and 1 meaning that the two areas overlap and have equal dimensions. We use IoU to compare the foreground regions of the class activation maps with the ground-truth bounding boxes. The computation of the class activation maps and of the metrics does not depend on the number of Saints in the painting, because every Iconclass category is associated with a different activation map independent of the others. All the reported results are valid regardless of the number of Saints.

#### 4.5. Quantitative Analysis

This section presents the results of comparing quantitatively the effectiveness of the class activation maps algorithms in the localization of iconography classes and their symbols.

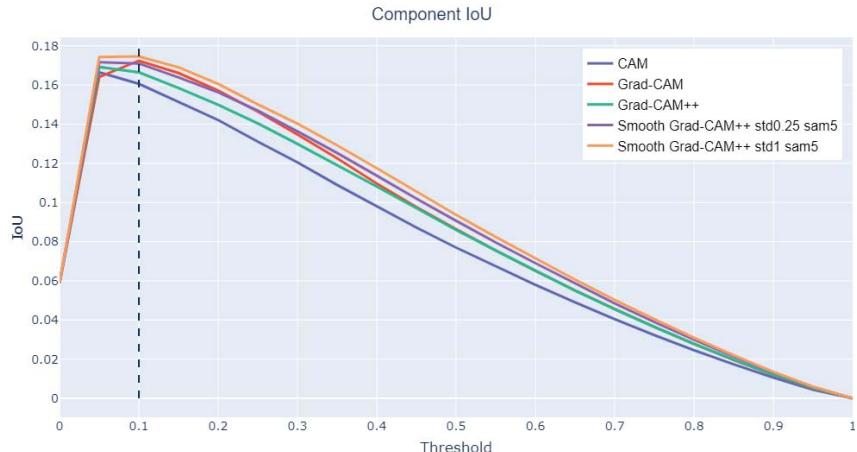
Smooth Grad-CAM++ is the only method that requires hyper-parameters: the standard deviation  $\sigma$  and the number of samples  $s$ . To set the hyper-parameter values, a grid-search was executed in the following space:  $\sigma \in \{0.25, 0.5, 1\}$  and  $s \in \{5, 10, 25\}$ . Only the best and worst Smooth Grad-CAM++ configurations are reported, to emphasize the boundary values reached by this algorithm. The number of samples is found to barely affect the results, whereas the standard deviation has more impact. To reduce the computational cost a lower number of samples is preferable.

##### Component IoU

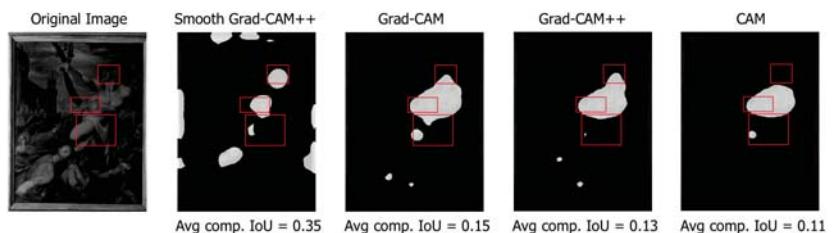
This metric evaluates how well the class activation map focuses on the individual Saints' symbols. First, the class activation map foreground area is divided into connected components, i.e., groups of pixels connected to each other. The IoU value is calculated between each ground-truth bounding box and the connected components that intersect it. Then, the average IoU across all symbol classes is taken. This procedure is repeated for all threshold values.

Figure 6 shows that the best results are obtained by Smooth Grad-CAM++ with a standard deviation  $\sigma = 1$  and a number of samples  $s = 5$ . The reason for this is that Smooth Grad-CAM++ tends to produce smaller and more focused areas, which yield more connected components and better coverage of the distinct symbols. Grad-CAM tends to create larger and more connected areas. This increases the size of the union and such an increase is not compensated by an equivalent increase of the intersection, which motivates

the lower IoU values. In all the considered class activation maps variants, the component IoU peak is found for a threshold value  $t \in \{0.05, 0.1\}$ . Grad-CAM creates larger and more connected regions, and thus, a higher threshold is needed to obtain the same number of components as the other methods. This explains why the component IoU peak is found at a higher threshold. Figure 7 (San Sebastian’s Martyrdom, Giovanni Maria Butteri, 1550–1559) compares the component IoU values produced on a sample image by different class activation maps algorithms. For the same threshold value, Smooth Grad-CAM++ creates more and better focused components.



**Figure 6.** Component IoU at varying threshold levels.



**Figure 7.** Different values of component IoU produced by different class activation map algorithms (Smooth Grad-CAM++ with  $\sigma = 1$  and  $s = 5$ ) at threshold  $t = 0.1$ . Ground-truth bounding boxes are shown in red.

#### Global IoU

An alternative metric is the IoU between the union of all the bounding boxes in the image and the entire foreground area of the class activation map taken at a given threshold. This metric is calculated for all threshold values and assesses how the class activation map focuses on the whole representation of the Saint, favoring those class activation maps methods that generate wider and more connected areas rather than separated components. Figure 8 shows that Grad-CAM is significantly better than the other analyzed methods. As already observed, Grad-CAM tends to spread over the entire figure and covers better the Saint and the associated symbols. Due to the complementary role of the component and global IoU metrics, the method with the best component IoU (Smooth Grad-CAM++ with  $\sigma = 1$  and  $s = 5$ ) has the worst global IoU. Differently from the component IoU, the global IoU peak position on the  $x$  axis does not change across methods, because the influence of the number of components is less relevant when the global metric is computed.

Figure 9 (Saint Jerome in the study, nd, 1604) compares the global IoU values produced on a sample image by different class activation map algorithms. For the same threshold value, Grad-CAM generates wider areas that cover more foreground pixels.

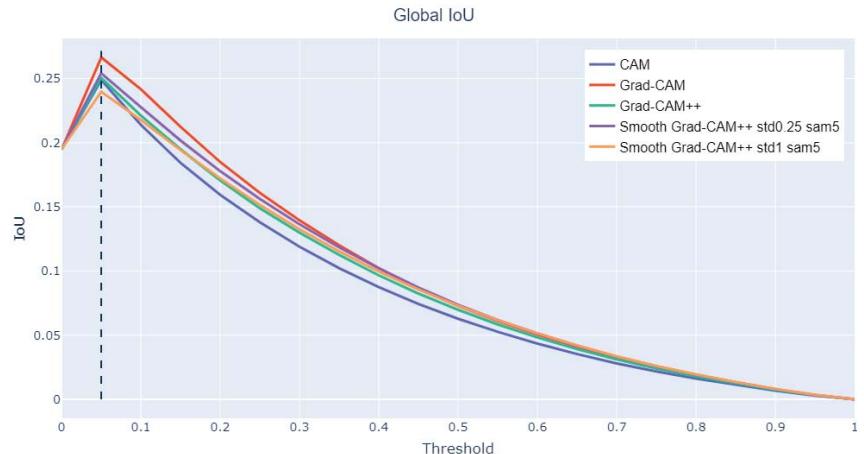


Figure 8. Global IoU at varying threshold levels.

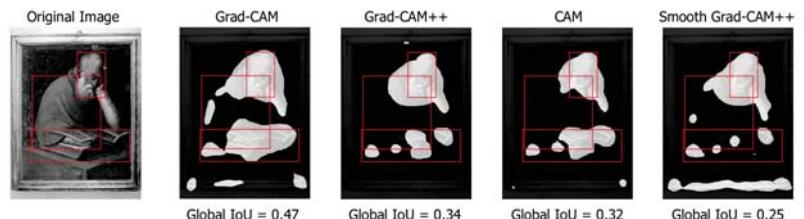


Figure 9. Different values of global IoU produced by different class activation map algorithms (Smooth Grad-CAM++ with  $\sigma = 1$  and  $s = 5$ ) at threshold  $t = 0.05$ . Manually annotated symbol bounding boxes are shown.

#### Bounding box coverage

When analyzing the class activation map algorithms, a factor to consider is also how many bounding boxes are covered by each class activation map. This metric alone is not enough to characterize the performance because a trivial class activation map that covers the entire image would have 100% coverage. However, coupled with the two previous metrics, it can give information about which method is able to generate class activation maps that can highlight a large fraction of the iconographic symbols that an expert would recognize. The bounding box coverage metric considers that a bounding box is covered by the class activation map only if their intersection is greater than or equal to 20% of the bounding box area. Figure 10 presents the results: Grad-CAM and Smooth Grad-CAM++ intersect, on average, more bounding boxes than the other methods. This result confirms that Grad-CAM covers wider areas, while focusing on the correct details at the same time. The worst method, CAM, performs poorly also in the two previous metrics. This indicates that it generates class activation maps that are smaller and less focused on the iconographic symbols with respect to the other approaches.

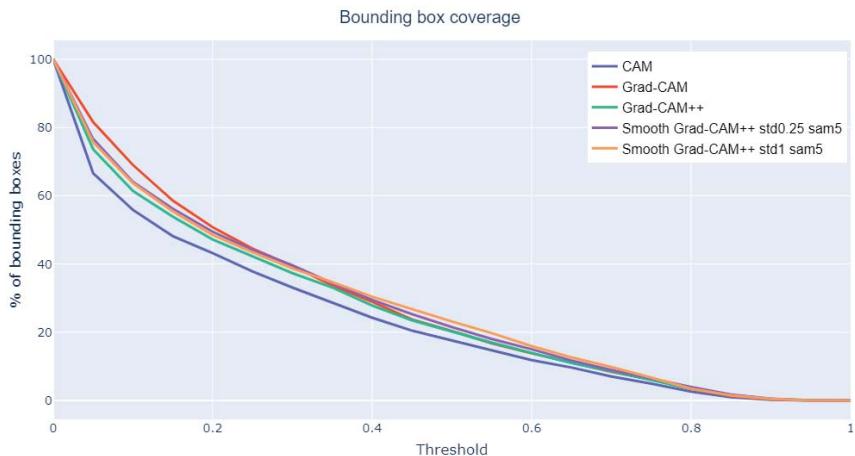


Figure 10. Bounding box coverage at varying threshold values.

#### Irrelevant attention

When evaluating the global IoU, a low value can occur for two reasons: (1) the two areas have a very small intersection or (2) the two areas overlap well but one is much larger than the other. Thus, an analysis on how much the class activation maps focus on irrelevant parts of the image helps characterizing low global IoU values. Irrelevant attention corresponds to the percentage of class activation map area outside any bounding box. Figure 11 shows that CAM has the less irrelevant attention, coherently with the previous results. Figure 12 (Madonna with Child and Infant St. John surrounded by Angels, Tiziano Vecellio, 1550) compares the irrelevant attention values produced on a sample image by different class activation map algorithms. For the same threshold value, CAM generates smaller irrelevant areas whereas Grad-CAM and Smooth-Grad-CAM++ include more irrelevant regions corresponding to the painting frame. The tendency of Smooth Grad-CAM++ to focus on irrelevant areas can be seen also in Figures 7 and 9.

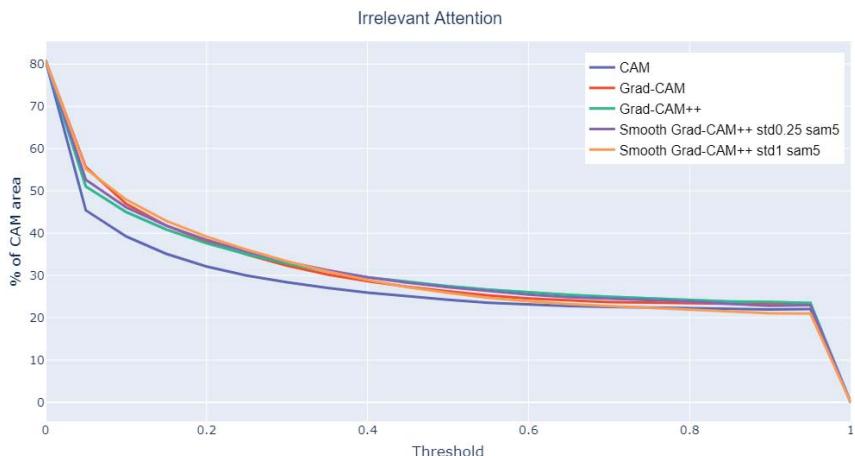
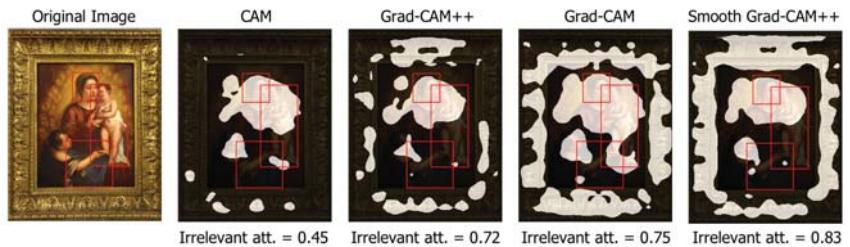


Figure 11. Irrelevant attention at varying threshold values.



**Figure 12.** Different values of irrelevant attention produced by different class activation map algorithms (Smooth Grad-CAM++ with  $\sigma = 1$  and  $s = 5$ ) at threshold  $t = 0.1$ . Manually annotated symbol bounding boxes are reported.

#### 4.6. Qualitative Analysis

This section presents a qualitative analysis of the results obtained by the different class activation map algorithms highlighting their capabilities and limitations. Each example shows the original image, the class activation maps generated by each algorithm (with background in black and foreground in white) and the ground-truth bounding boxes.

##### Positive examples

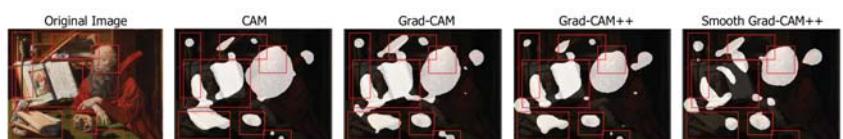
Figure 13 (Saint Jerome in his Study, Jan van Remmerswale, 1533) shows an example in which all the algorithms focus well on the iconographic symbols. The image contains seven symbols with different size, shape and position, which are all identified and separated by the class activation map algorithms. The irrelevant area on the top right corresponds to a piece of the cardinal's vest that has the same color and approximate shape of the cardinal's galero appearing in many paintings of Saint Jerome.

Figure 14 (St. Peter, Antonio Veneziano, 1369–1375) shows an example in which all the algorithms perform well on a painting in which the visibility of the symbols is very low. All class activation map algorithms identify four out of the five symbols. The central ground-truth bounding box is not identified because it corresponds to a rather generic attribute (the bishop's vest), which is not evident in the drawing. Only CAM misses the book, which the other algorithms identify by focusing on the characteristic marks on the spine of the book or on the lock. The example of Figure 14 and many similar ones of black and white and poor quality images highlight the ability of class activation map algorithms to extract useful maps also when the image has low discriminative features.

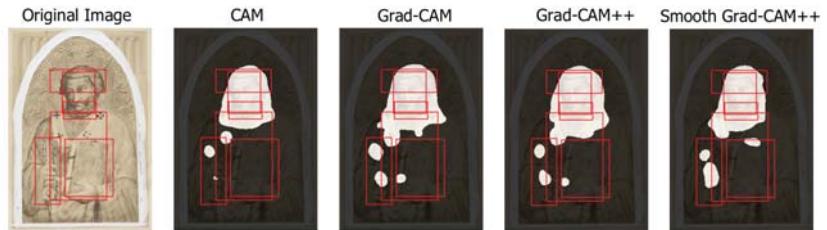
A counterexample of the difficulty of detecting such generic attributes as the vest is illustrated in Figure 15 (Saint Dominic, Carlo Crivelli, 1472). The vest is identified thanks to a specific detail: the change of color typical of the black and white Dominican habit.

##### Negative examples

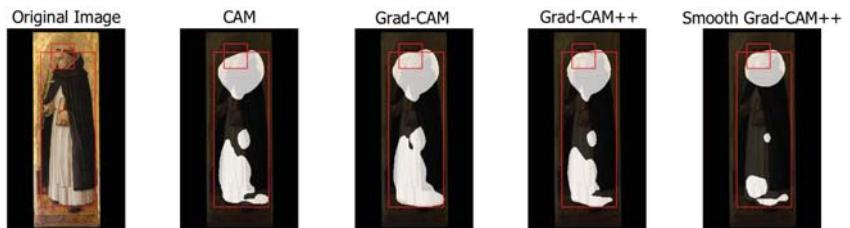
Class activation maps algorithms tend to fail consistently in two cases: when multiple symbols are too close or have a substantial overlap and when the representation of a symbol is rather generic and covers a wide area of the image. Figure 16 (Penitent St. Peter, Jusepe de Ribera, 1600–1649) illustrates a typical example: Saint Peter's bushy hair and beard are merged into a single region and the vest, which is a rather generic attribute, is missed completely or highlighted only through small irrelevant details.



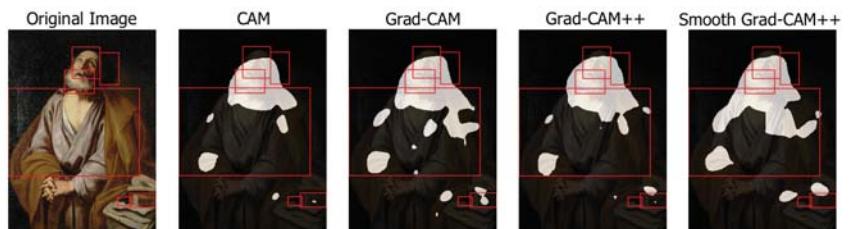
**Figure 13.** Class activation maps with seven recognized symbols associated with Saint Jerome.



**Figure 14.** Class activation maps extracted from a drawing of Saint Peter. Four out of five symbols are identified despite their low visibility.



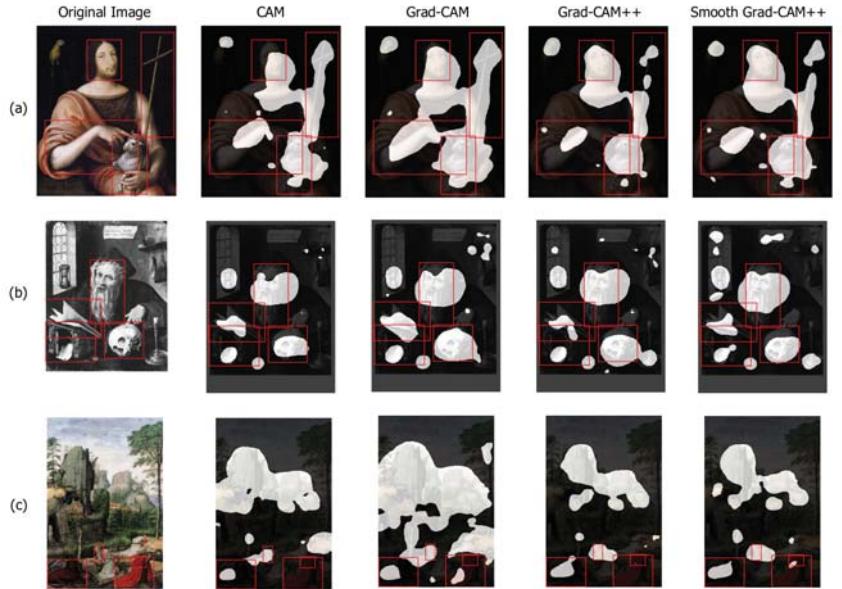
**Figure 15.** Class activation maps extracted from a painting of Saint Dominic. The rather generic vest attribute is identified by focusing on its double color.



**Figure 16.** Class activation maps with merged symbols and missed generic attributes.

#### Relevant irrelevant regions

An interesting case occurs when the class activation map algorithms focus on an apparently irrelevant area, which instead contains a relevant iconographic attribute not present in the ground truth. Figure 17 illustrates three examples. The painting of Saint John the Baptist (a) (portrait of François I as St John the Baptist, Jean Clouet, 1518) contains an apparently irrelevant area in the top left, which focuses on a bird. This is a less frequent attribute of the Saint that is not listed in the iconographic symbols used to annotate the images but appears in some of the paintings. The same happens with Saint Jerome (b) (Saint Jerome, Albrecht Durer, 1521), where the class activation map algorithms highlight an hourglass, an infrequent symbol present only in a subset of the ArtDL images and not used in the annotation. Finally, another case occurs with the iconography of Saint Jerome (c) (Landscape with St. Jerome, Simon Bening, 1515–1520), where the class activation map algorithms focus on the outdoor environment. This is a well-known symbol associated with the Saint, who retired in the wilderness, but one that is hard to annotate with bounding boxes and thus purposely excluded from the ground truth.



**Figure 17.** Class activation maps highlighting regions containing relevant iconographic attributes not present in the ground truth: a bird associated with Saint John the Baptist (a) an hourglass associated with Saint Jerome (b) and the wilderness where Saint Jerome retired (c).

#### Confusion with unknown co-occurring class

Figure 18 (Baptism of Christ, Pietro Perugino, 1510) presents an example in which all analyzed variants make confusion between Saint John the Baptist and Jesus Christ. The latter is an Iconclass category too, but not one represented in the ArtDL dataset. Given the prevalence of paintings depicting Saint John the Baptist in the act of baptizing Christ over those where the Saint occurs alone, the CAM output highlights both the figures. This ambiguity would reduce if the dataset were annotated with the Iconclass category for Jesus.



**Figure 18.** Class activation maps with confusion between Saint John the Baptist and Jesus Christ.

#### Bounding Box Generation

The goal of the presented work is to compare the effectiveness of alternative class activation map algorithms in isolating the salient regions of artwork images that have the greatest impact for the attribution of a specific iconography class. The capacity of a class activation map algorithm to identify precisely the areas of an image that correspond to the whole Saint or to one of the iconographic symbols that characterize him/her can help build a training set for the object detection task. The class activation map can be used as a replacement of the manual annotations necessary for creating a detection training set by computing the smallest bounding boxes that comprise the foreground area and

using such automatically generated annotations for training an object detector. This approach is known as weakly supervised object detection and is an active research area [48]. To investigate the potential of the class activation maps to support weakly supervised object detection, the region proposals obtained by drawing bounding boxes around the connected components of the class activation maps have been compared visually with the ground-truth bounding boxes of the iconographic symbols. For completeness, we have also computed the bounding boxes surrounding all the foreground pixels and compared them with manually created bounding boxes surrounding the whole Saints. The candidate region proposals to use as automatic bounding boxes have been identified with the following heuristic procedure.

1. Collect the images on which all the four methods satisfy a minimum quality criterion: for symbol bounding boxes component IoU greater than 0.165 at threshold 0.1 (see Figure 6) and for whole Saint bounding boxes global IoU greater than 0.24 at threshold 0.05 (see Figure 8);
2. Compute the Grad-CAM class activation map of the selected images and apply the corresponding threshold: 0.1 for symbol bounding boxes and 0.05 for whole Saint bounding boxes;
3. Only for symbol boxes: split the class activation maps into connected components. Remove the components whose average activation value is less than half of the average activation value of all components. This step filters out all the foreground pixels with low activation that usually correspond to irrelevant areas (Figure 12);
4. For each Iconclass category, draw one bounding box surrounding each component (symbol bounding boxes) and one bounding box surrounding the entire class activation map (whole Saint bounding boxes).

In the procedure above, Grad-CAM is chosen to compute the candidate symbol and whole Saint bounding boxes, because it has the highest value of the bounding box coverage metrics (together with Smooth Grad-CAM++) and covers wider areas, at the same time, focusing on the correct details.

#### Symbol bounding boxes

Figure 19 presents some examples of the computed symbol bounding boxes (green) compared with the ground-truth bounding boxes (red). The proposed procedure is able to generate boxes that in many cases correctly highlight and distinguish the most important iconographic symbols present in the images. When the symbols are grouped in a small area (e.g., the bushy hair and beard of Saint Peter), the procedure tends to generate one component that covers all of them, thus creating only one bounding box. Sometimes, elements in the image that have not been manually annotated in the ground truth are correctly detected (e.g., the scroll in the hand of Saint John the Baptist in the first painting of Figure 19).

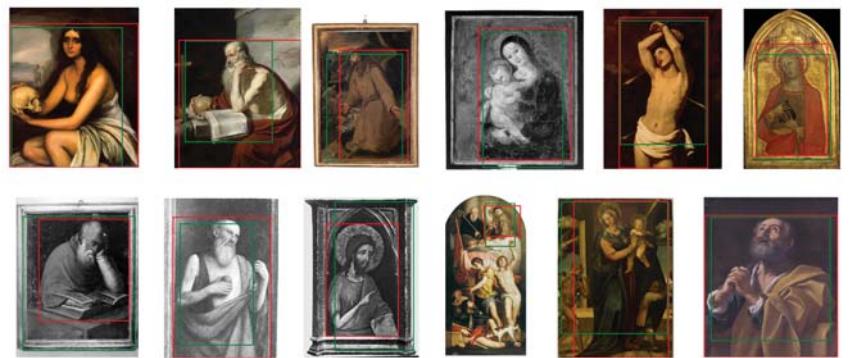
#### Whole Saint bounding boxes

Figure 20 illustrates some examples of computed whole Saint bounding boxes (green) compared with the ground-truth boxes (red). The automatically generated bounding boxes localize almost entirely the Saint’s figure and include only very small irrelevant areas.

Figures 19 and 20 show that the simple procedure for processing class activation map outputs is sufficient to generate good quality bounding boxes that can act as a proxy to the ground truth for training a fully supervised object detector.



**Figure 19.** Examples of symbols bounding boxes generated from Grad-CAM (green) and manually annotated (red).



**Figure 20.** Examples of Saints bounding boxes generated from Grad-CAM (green) and manually annotated (red).

#### Quantitative evaluation of whole Saint bounding boxes

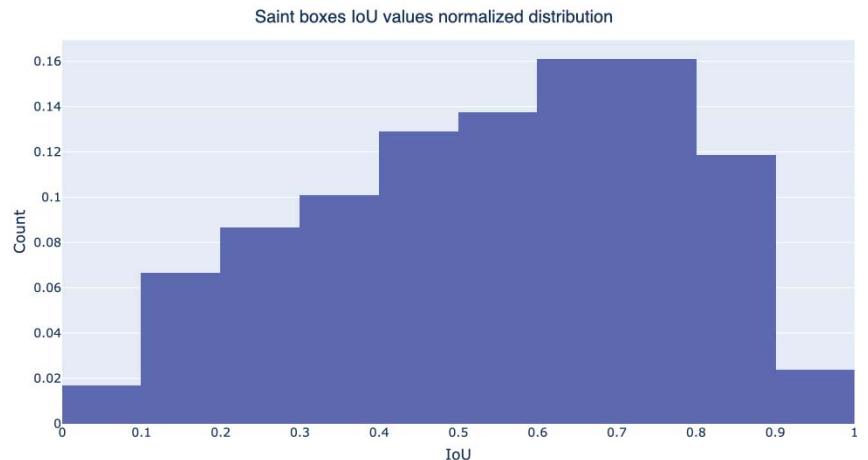
For the whole Saint case, each estimated bounding box can be labeled with the iconography class of the corresponding Saint portrayed in the image. In this way, it is possible to quantify the coincidence between the bounding box of the ground truth and the bounding box computed from the class activation map. For this purpose, three object detection metrics have been computed: the average IoU value between the GT and the estimated bounding boxes, mean Average Precision and GT-known Loc. The latter is used in several works ([49–51]) to evaluate the localization accuracy of object detectors and is defined as the percentage of correct bounding boxes. A bounding box is considered correct only when the IoU between the GT box (for a specific class) and the estimated box (for the same class) is greater than 0.5. Results are reported in Table 3: Grad-CAM confirms as the method with the best performances, Smooth-Grad-CAM++ yields similar results, and CAM is the worst performing method in all the computed metrics. Grad-CAM produces bounding boxes that on average have 0.55 IoU with the GT boxes and the GT-known Loc metric shows that ~61% of those boxes have an IoU value greater than 0.5. Figure 21 presents the normalized distribution of IoU values for Grad-CAM. We can observe that ~83% of the generated boxes have an IoU value greater than 0.3 and that most values are in the range between 0.4 and 0.9, with ~12% having an IoU greater than 0.9. Table 4 shows the mAP values obtained with GradCAM on the ten ArtDL classes.

The whole Saint estimated bounding boxes appear to be suitable for creating the pseudo ground truth for training an object detector with the weakly supervised approach.

Two observations motivate the viability of Grad-CAM for this purpose. As in the GT-known Loc metrics, the goodness of an object detection is usually evaluated with a minimal IoU threshold of 0.5 and the boxes generated automatically with Grad-CAM obtain 0.55 IoU on average, which suggests that the automatically estimated bounding boxes have a quality similar to the bounding boxes produced by a fully supervised object detector, albeit inferior to the quality of the bounding boxes created by humans. Grad-CAM, which is designed to be an interpretability technique, can be used also to estimate bounding boxes that reach 31.6% mAP on cultural heritage data without any optimization. This finding compares well with the fact that methods designed and optimized specifically for weakly supervised object detection reach values around 14% on artworks datasets similar to ArtDL [10,52]. For this reason, simple and generic techniques such as Grad-CAM, which can localize multiple Saint instances and even multiple characteristic features, are a promising starting point for advancing weakly supervised object detection studies in the cultural heritage domain.

**Table 3.** Average IoU, GT-Known accuracy and mAP values for the whole Saint bounding boxes estimated with the four analyzed class activation map techniques. The values are calculated with an activation threshold equal to 0.05.

Method	Average IoU	GT-Known Loc (%)	mAP (at IoU $\geq 0.5$ )
CAM	0.489	49.70	0.206
GradCAM	0.551	61.20	0.316
GradCAM++	0.529	59.88	0.292
Smooth-GradCAM++	0.544	61.18	0.307



**Figure 21.** Normalized distribution of IoU values between whole-Saint Grad-CAM estimated bounding boxes and ground-truth bounding boxes.

**Table 4.** Mean Average Precision (mAP) values for each class of the ArtDL dataset. Bounding boxes are estimated with GradCAM.

Anthony	John	Paul	Francis	Magdalene	Jerome	Dominic	Virgin	Peter	Sebastian
0.076	0.289	0.173	0.33	0.616	0.228	0.142	0.442	0.399	0.468

## 5. Conclusions and Future Work

This work has presented a comparative study about the effectiveness of class activation maps as a tool for explaining of how a CNN-based classifier recognizes the Iconclass categories present in images portraying Christian Saints. The symbols relevant to the identification of the Saints were annotated with bounding boxes and the output of the class activation maps algorithms were compared to the ground truth using four metrics. The analysis shows that Grad-CAM achieves better results in terms of global IoU and covered bounding boxes and Smooth Grad-CAM++ scores best in the component IoU thanks to its precision in delineating individual small size symbols. The irrelevant attention metric promotes the original CAM algorithm as the best approach, but the low component IoU and box coverage complement such an evaluation showing that CAM covers too small areas. While for natural images Smooth Grad-CAM++ outperforms the other three algorithms [18], in our use case Grad-CAM is the method of choice for deriving the bounding boxes from class activation maps necessary to train a weakly supervised detector.

Future work will concentrate on the comparison of other activation mapping techniques [50,51,53,54]. In particular, [50,51] are based on the re-training of the network, an approach quite different from the currently analyzed alternatives. The results of the CAMs algorithms selection will be used to pursue the ultimate goal of our research, which is to use the output of class activation maps to create training datasets for weakly supervised iconographic symbol detection and segmentation. The implementation of an automated system for iconographic analysis of artworks could promote the development of educational applications for art history experts and students. Finally, another future research path consists in addressing more complex Iconclass categories involving complex scenes (e.g., the crucifixion, the nativity, the visitation of the magi, etc.) and in exploring the iconography of other cultures.

**Author Contributions:** Conceptualization, N.O.P.V., F.M., P.F. and R.d.S.T.; Methodology, N.O.P.V., F.M., P.F. and R.d.S.T.; Validation, N.O.P.V., F.M., P.F. and R.d.S.T.; Writing—original draft, N.O.P.V., F.M., P.F. and R.d.S.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. This data can be found here: <http://www.artdl.org> (accessed on 29 June 2021).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Panofsky, E. *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*; Routledge Taylor and Francis Group: New York, NY, USA, 1939; p. 262.
2. Proulx, D.A. *A sourcebook of Nasca Ceramic Iconography: Reading a Culture through Its Art*; University of Iowa Press: Iowa City, IA, USA, 2009.
3. Parani, M.G. *Reconstructing the Reality of Images: Byzantine Material Culture and Religious Iconography 11th-15th Centuries*; Brill: Leiden, The Netherlands, 2003; Volume 41.
4. Van Leeuwen, T.; Jewitt, C. *The Handbook of Visual Analysis*; Sage: Thousand Oaks, CA, USA, 2001; pp. 100–102.
5. King, J.N. *Tudor Royal Iconography: Literature and Art in an Age of Religious Crisis*; Princeton University Press: Princeton, NJ, USA, 1989.
6. Roberts, H.E. *Encyclopedia of Comparative Iconography: Themes Depicted in Works of Art*; Routledge: London, UK, 2013.
7. Zujovic, J.; Gandy, L.; Friedman, S.; Pardo, B.; Pappas, T.N. Classifying paintings by artistic genre: An analysis of features classifiers. In Proceedings of the 2009 IEEE International Workshop on Multimedia Signal Processing, Rio de Janeiro, Brazil, 5–7 October 2009; pp. 1–5.
8. Shamir, L.; Tarakhovsky, J.A. Computer Analysis of Art. *J. Comput. Cult. Herit.* **2012**, *5*. [[CrossRef](#)]
9. Cai, H.; Wu, Q.; Corradi, T.; Hall, P. The Cross-Depiction Problem: Computer Vision Algorithms for Recognising Objects in Artwork and in Photographs. *arXiv* **2015**, arXiv:1505.00110.

10. Gonthier, N.; Gousseau, Y.; Ladjal, S.; Bonfait, O. Weakly supervised object detection in artworks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
11. Milani, F.; Fraternali, P. A Data Set and a Convolutional Model for Iconography Classification in Paintings. *arXiv* **2020**, arXiv:2010.11697.
12. Sun, K.H.; Huh, H.; Tama, B.A.; Lee, S.Y.; Jung, J.H.; Lee, S. Vision-Based Fault Diagnostics Using Explainable Deep Learning With Class Activation Maps. *IEEE Access* **2020**, *8*, 129169–129179. [[CrossRef](#)]
13. Patro, B.; Lunayach, M.; Patel, S.; Namboodiri, V. U-CAM: Visual Explanation Using Uncertainty Based Class Activation Maps. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 7443–7452. [[CrossRef](#)]
14. Yang, S.; Kim, Y.; Kim, Y.; Kim, C. Combinational Class Activation Maps for Weakly Supervised Object Localization. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 2930–2938. [[CrossRef](#)]
15. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
16. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
17. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018. [[CrossRef](#)]
18. Omeiza, D.; Speakman, S.; Cintas, C.; Weldermariam, K. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. *arXiv* **2019**, arXiv:1908.01224.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
20. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009; IEEE Computer Society: New York, NY, USA, 2009; pp. 248–255. [[CrossRef](#)]
21. Karayev, S.; Trentacoste, M.; Han, H.; Agarwala, A.; Darrell, T.; Hertzmann, A.; Winnemoeller, H. Recognizing image style. *arXiv* **2013**, arXiv:1311.3715.
22. Crowley, E.J.; Zisserman, A. *The State of the Art: Object Retrieval in Paintings Using Discriminative Regions*; British Machine Vision Association: Durham, UK, 2014.
23. Khan, F.S.; Beigpour, S.; Van de Weijer, J.; Felsberg, M. Painting-91: A large scale database for computational painting categorization. *Mach. Vis. Appl.* **2014**, *25*, 1385–1397. [[CrossRef](#)]
24. Strezoski, G.; Worring, M. Omniart: Multi-task deep learning for artistic data analysis. *arXiv* **2017**, arXiv:1708.00684.
25. Mao, H.; Cheung, M.; She, J. Deepart: Learning joint representations of visual arts. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1183–1191.
26. Bianco, S.; Mazzini, D.; Napoletano, P.; Schettini, R. Multitask painting categorization by deep multibranch neural network. *Expert Syst. Appl.* **2019**, *135*, 90–101. [[CrossRef](#)]
27. Castellano, G.; Vessio, G. Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. In *Neural Computing and Applications*; Springer: New York, NY, USA, 2021; pp. 1–20.
28. Santos, I.; Castro, L.; Rodriguez-Fernandez, N.; Torrente-Patino, A.; Carballal, A. Artificial Neural Networks and Deep Learning in the Visual Arts: A review. In *Neural Computing and Applications*; Springer: New York, NY, USA, 2021; pp. 1–37.
29. Zhao, W.; Zhou, D.; Qiu, X.; Jiang, W. Compare the performance of the models in art classification. *PLoS ONE* **2021**, *16*, e0248414. [[CrossRef](#)]
30. Gao, Z.; Shan, M.; Li, Q. Adaptive sparse representation for analyzing artistic style of paintings. *J. Comput. Cult. Herit. (JOCCH)* **2015**, *8*, 1–15. [[CrossRef](#)]
31. Elgammal, A.; Kang, Y.; Den Leeuw, M. Picasso, matisse, or a fake? Automated analysis of drawings at the stroke level for attribution and authentication. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
32. Crowley, E.J.; Zisserman, A. Of gods and goats: Weakly supervised learning of figurative art. *Learning* **2013**, *8*, 14.
33. Shen, X.; Efros, A.A.; Aubry, M. Discovering Visual Patterns in Art Collections with Spatially-consistent Feature Learning. *arXiv* **2019**, arXiv:1903.02678.
34. Kadish, D.; Risi, S.; Løvlie, A.S. Improving Object Detection in Art Images Using Only Style Transfer. *arXiv* **2021**, arXiv:2102.06529.
35. Banar, N.; Daelemans, W.; Kestemont, M. *Multi-Modal Label Retrieval for the Visual Arts: The Case of Iconclass*; Scitepress: Setúbal, Portugal, 2021.
36. Gonthier, N.; Gousseau, Y.; Ladjal, S. An analysis of the transfer learning of convolutional neural networks for artistic images. *arXiv* **2020**, arXiv:2011.02727.

37. Cömert, C.; Özbayoğlu, M.; Kasnakoglu, C. Painter Prediction from Artworks with Transfer Learning. In Proceedings of the IEEE 2021 7th International Conference on Mechatronics and Robotics Engineering (ICMRE), Budapest, Hungary, 3–5 February 2021; pp. 204–208.
38. Bellhi, A.; Ahmed, H.O.; Alfaqheri, T.; Bouras, A.; Sadka, A.H.; Foufou, S. Study and Evaluation of Pre-trained CNN Networks for Cultural Heritage Image Classification. In *Data Analytics for Cultural Heritage: Current Trends and Concepts*; Springer: Cham, Switzerland, 2021; p. 47.
39. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–42. [[CrossRef](#)]
40. Buhrmester, V.; Münch, D.; Arens, M. Analysis of explainers of black box deep neural networks for computer vision: A survey. *arXiv* **2019**, arXiv:1911.12116.
41. Gupta, V.; Demirer, M.; Bigelow, M.; Yu, S.M.; Yu, J.S.; Prevedello, L.M.; White, R.D.; Erdal, B.S. Using Transfer Learning and Class Activation Maps Supporting Detection and Localization of Femoral Fractures on Anteroposterior Radiographs. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; pp. 1526–1529.
42. Zhang, M.; Zhou, Y.; Zhao, J.; Man, Y.; Liu, B.; Yao, R. A survey of semi-and weakly supervised semantic segmentation of images. *Artif. Intell. Rev.* **2020**, *53*, 4259–4288. [[CrossRef](#)]
43. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
44. Qiu, S. Global Weighted Average Pooling Bridges Pixel-level Localization and Image-level Classification. *arXiv* **2018**, arXiv:1809.08264.
45. Lanzi, F.; Lanzi, G. *Saints and Their Symbols: Recognizing Saints in Art and in Popular Images*; Liturgical Press: Collegeville, MN, USA, 2004; pp. 327–342.
46. Wikipedia: Saint Symbolism. [https://en.wikipedia.org/wiki/Saint\\_symbolism](https://en.wikipedia.org/wiki/Saint_symbolism) (accessed on 24 April 2021).
47. Couplie, L.D. Iconclass: An iconographic classification system. *Art Libr. J.* **1983**, *8*, 32–49. [[CrossRef](#)]
48. Zhang, D.; Han, J.; Cheng, G.; Yang, M.H. Weakly Supervised Object Localization and Detection: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [[CrossRef](#)]
49. Singh, K.K.; Lee, Y.J. Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization. *arXiv* **2017**, arXiv:1704.04232.
50. Choe, J.; Shim, H. Attention-Based Dropout Layer for Weakly Supervised Object Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 2219–2228. [[CrossRef](#)]
51. Bae, W.; Noh, J.; Kim, G. Rethinking Class Activation Mapping for Weakly Supervised Object Localization. In *Part XV, Proceedings of the Computer Vision - ECCV 2020—16th European Conference, Glasgow, UK, 23–28 August 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J., Eds.; Lecture Notes in Computer Science; Springer: New York, NY, USA, 2020; Volume 12360, pp. 618–634. [[CrossRef](#)]
52. Gonthier, N.; Ladjal, S.; Gousseau, Y. Multiple instance learning on deep features for weakly supervised object detection with extreme domain shifts. *arXiv* **2020**, arXiv:2008.01178.
53. Wang, H.; Du, M.; Yang, F.; Zhang, Z. Score-cam: Improved visual explanations via score-weighted class activation mapping. *arXiv* **2019**, arXiv:1910.01279.
54. Zhao, G.; Zhou, B.; Wang, K.; Jiang, R.; Xu, M. Respond-cam: Analyzing deep models for 3d imaging data by visualizations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: New York, NY, USA, 2018; pp. 485–492.



Article

# Classification of Geometric Forms in Mosaics Using Deep Neural Network

Mridul Ghosh <sup>1,2</sup>, Sk Md Obaidullah <sup>2</sup>, Francesco Gherardini <sup>3,\*</sup> and Maria Zdimalova <sup>4</sup>

<sup>1</sup> Department of Computer Science, Shyampur Siddheswari Mahavidyalaya, Howrah 711312, India; mridulxyz@gmail.com

<sup>2</sup> Department of Computer Science & Engineering, Aliah University, Kolkata 700160, India; sk.obaidullah@aliah.ac.in

<sup>3</sup> Department of Engineering “Enzo Ferrari”, University of Modena and Reggio Emilia, 41121 Modena, Italy

<sup>4</sup> Department of Mathematics and Descriptive Geometry, Slovak University of Technology in Bratislava, 810 05 Bratislava, Slovakia; maria.zdimalova@stuba.sk

\* Correspondence: francesco.gherardini@unimore.it



**Citation:** Ghosh, M.; Obaidullah, S.M.; Gherardini, F.; Zdimalova, M. Classification of Geometric Forms in Mosaics Using Deep Neural Network. *J. Imaging* **2021**, *7*, 149. <https://doi.org/10.3390/jimaging7080149>

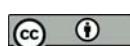
Academic Editors: Gennaro Vessio, Giovanna Castellano and Fabio Bellavia

Received: 9 July 2021

Accepted: 15 August 2021

Published: 18 August 2021

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** The paper addresses an image processing problem in the field of fine arts. In particular, a deep learning-based technique to classify geometric forms of artworks, such as paintings and mosaics, is presented. We proposed and tested a convolutional neural network (CNN)-based framework that autonomously quantifies the feature map and classifies it. Convolution, pooling and dense layers are three distinct categories of levels that generate attributes from the dataset images by introducing certain specified filters. As a case study, a Roman mosaic is considered, which is digitally reconstructed by close-range photogrammetry based on standard photos. During the digital transformation from a 2D perspective view of the mosaic into an orthophoto, each photo is rectified (i.e., it is an orthogonal projection of the real photo on the plane of the mosaic). Image samples of the geometric forms, e.g., triangles, squares, circles, octagons and leaves, even if they are partially deformed, were extracted from both the original and the rectified photos and originated the dataset for testing the CNN-based approach. The proposed method has proved to be robust enough to analyze the mosaic geometric forms, with an accuracy higher than 97%. Furthermore, the performance of the proposed method was compared with standard deep learning frameworks. Due to the promising results, this method can be applied to many other pattern identification problems related to artworks.

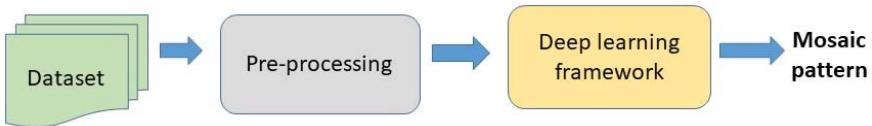
**Keywords:** deep learning algorithm; convolutional neural networks; pattern classification; image-based reconstruction; cultural heritage

## 1. Introduction

The application of science and engineering to the analysis of artifacts and artworks such as paintings, mosaics and statues dates back several centuries [1–3]. However, only over the past few decades have the analytical methods developed in the mathematical, IT and physical sciences been able to gather information from the past and contribute to the analysis, interpretation and dissemination in the fine arts. In the past, there was a historical division between science and the humanities, so the interaction between these two fields has never been natural. For example, the application of signal and image processing techniques for the analysis and restoration of artworks was a very uncommon practice. Lately, there has been a greater and growing attention and interest in processing image data of artworks for storage, transmission, representation and analysis, and an increasing number of scientists with a background in analytical and mathematical techniques has approached this field, in an interdisciplinary way. There are several ways in which image processing can find significant applications in the fields of fine arts and cultural heritage. Among them, three main areas of application can be identified: obtaining a digital version of

traditional photographic reproductions, pursuing imaging diagnostics and implementing virtual restoration [1,2,4]. Obtaining the exact reproduction and explanation of an artwork was one of the first developments in the first area, which includes the process of archiving, retrieving and disseminating data and derives all the benefits from the digital format [1–6]. In the second area of imaging diagnostics, digital images are used to detect and document the state of preservation of artifacts [7], as in the case of the noninvasive techniques based on imaging in different spectral regions used for the investigation of paintings [8]. In the third area, the image processing techniques can be used as a guide to the actual restoration of fine arts (computer-guided restoration), or they can produce a digitally restored version of the artwork. In some activities, the computer is more suitable than traditional artistic tools. Examples of such activities are filtering, geometric transformation of an image, segmentation and pattern recognition. Using digital technologies, every change to the image can be seen on the screen almost in real time. Moreover, images and data can be edited, filtered and processed with minimal material costs even when complicated operations are performed, e.g., changes in colors, brightness or contrast [5,9–12]. A further development consists of applying computer vision, an area of artificial intelligence, to recognize patterns of the historical art heritage [6,13].

In this scenario, this paper presents a method to perform the recognition of geometrical patterns in fine arts, thanks to image processing techniques. In particular, we developed and tested a deep learning-based framework to classify the geometric forms and patterns of floor mosaics, which consist of an arrangement of tiles usually characterized by jagged and undefined boundaries or surface irregularities. The workflow of the proposed method is shown in Figure 1.



**Figure 1.** The workflow of the proposed method.

The paper is organized as follows: In Section 2, we introduce methods of image processing applied to fine arts, involving machine learning and deep learning-based techniques. Section 3 describes the proposed method based on deep neural networks. Section 4 introduces the case study. Section 5 presents the experiments resulting from the application of the deep neural network framework to the dataset and the results achieved. In Section 6, some final remarks and open questions close the paper.

## 2. Related Work

This section proposes a literature survey dealing with various methods of image processing applied to fine arts, involving machine learning and deep learning-based techniques. In [14–18], image processing techniques for art investigation are applied to the detection of defects and cracks, as well as to the removal of defects and canvas from high-resolution acquisition of paintings. Examples of these kinds of methods include the use of sparse representations and the removal of cradling artifacts in X-ray images of panel paintings [15] and the automated crack detection using the Ghent Altarpiece [16], employed as guidance during its ongoing restoration.

Various methods of automatic image segmentation are used in the literature aiming at identifying regions in an image and labeling them as different classes. The main applications are pattern recognition for classifying paintings [19–23] or the authentication of fine arts (e.g., of paintings) [24]. These image segmentation methods include the following: The thresholding methods transform a grey-scale image into a binary image, where the algorithm evaluates the differences among neighboring pixels to find object boundaries [25–27]. The region growing methods are based on an expansion of an object detected inside of

an object [28,29] by selecting object seed pixels (inside an area to be detected) and then searching for neighboring pixels with similar intensities to the object seed pixels. In the level sets, the algorithm will converge at the boundary of the object where the differences are the highest. In the graph-cut method [30–32], firstly proposed by Wu and Leahy [30], each image is represented as a graph of nodes: each node corresponds to an image pixel, and links connecting the nodes are called edges; a pathway is constructed connecting all the edges to travel across the graph.

Aggregation methods are important as well for image resampling [33] or denoising [34]: When an appropriate scale or resolution is determined, the next step is to obtain the corresponding images. In the case of low scale or resolution, resampling techniques are often used to interpolate an image into a desired resolution, and aggregation is a particular resampling technique widely practiced for “up-scaling” image data from high resolution to low resolution [33].

This paper particularly focuses on deep learning [35,36], which is a kind of machine learning that uses several levels of neurons with complicated architectures or nonlinear changes to represent greater interpretations of information. With the growing volume of information and computing power, neural systems having increasingly sophisticated architecture have been of great interest and are used in a variety of disciplines. Some examples of applications in image processing and in fine arts are as follows: Image segmentation using a neural network has recently been used as a very strong tool for image processing [22,37]; recently, even convolutional neural networks have been applied to paintings [38]. In [39], a novel deep learning framework is developed to retrieve similar architectural floor plan layouts from a repository, analyzing the effect of individual deep convolutional neural network layers for the floor plan retrieval task. In [40] the results of a novel method for building structure extraction in urbanized aerial images are presented. Most of the methods are based on CNN. Similarly, in [41], the use of deep neural networks for object detection in floor plan images is investigated, evaluating the use of object detection architectures to recognize furniture objects, doors and windows in floor plans.

Gomez-Rios et al. [42] classified the textures of underwater coral patterns based on a CNN-based transfer learning-based approach. To work on diverse data and evaluate the performance of the proposed approach, they used data augmentation. The adoption of a deep neural network can significantly improve phase demodulation efficiency from a singular fringe sequence [43]. Their system was developed to anticipate several subsequent outcomes that may be used to calculate an incoming fringe pattern cycle. They collected fringe pictures of diverse situations to produce training input while the systems are being trained. The neural network blindly took only one input fringe sequence and produced the associated estimations of such transitional outcomes at great accuracy. Sandelin [44] proposed a Mask R-CNN-based technique for floor plan pictures and segmented the walls, windows, chambers and doors. This method showed good performance even in noisy images. Vilnrotter et al. [45] proposed a technique to generate appropriate naturalistic texture characteristics. The fundamental method of edge characteristics to determine an initial, incomplete identification of the components was discussed. The graphic components were extracted using such characterization. The components were classified into types and topological connections with them. The formulations were proven to be beneficial for texture identification and recurrent pattern restoration.

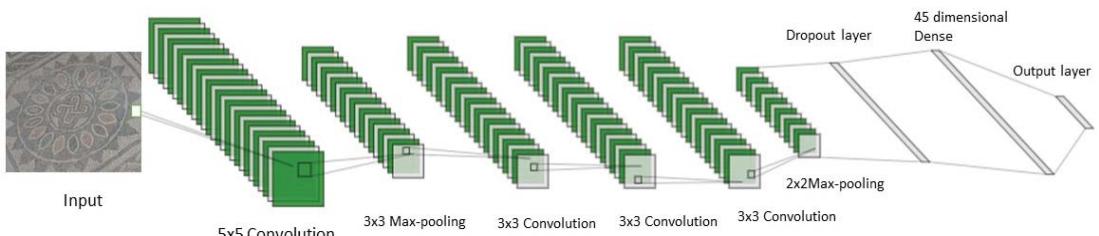
With a particular focus on mosaics, most of the related computer applications deal with their digital reconstruction using image-based techniques (i.e., photogrammetry) for documentation and analysis [46–49]. Besides, literature presents a few examples of image processing applications: In [50], a registration method in the framework of a restoration process of a medieval mosaic to compare a historical black and white photograph with a current digital one is presented. In [51], an algorithm that exploits deep learning and image segmentation techniques is presented to obtain a digital (vector) representation of a mosaic. In [52], the restoration of historical photographs of an ancient mosaic (by removing noise, deburring the image and increasing the contrast) and then the removal

of geometrical difference between images by means of the multimodal registration using mutual information is presented; the final identification of differences between the photos indicates the changes in the mosaic during the centuries. In [53], Falomir et al. presented a mathematical method for calculating a likeness score among qualitative assessments of item structure, color and dimension in digitized pictures. The closeness scores calculated are dependent on compositional cluster maps or intermediate distances, as per the specification of the subjective characteristics. The outcome using prior techniques was enhanced by using an estimated identification process among item characteristics of a tile mosaic assembly.

### 3. Proposed Method

In this paper, we propose a deep learning-based framework to classify the forms of fine arts, such as paintings and mosaics. The algorithm is able to classify the geometrical forms constituting the patterns, even if they are partially deformed. This deep learning [54] is a type of machine learning that eliminates the need for manual processing of features. Images are immediately fed into this system, and the final categorization is returned. Due to its high capacity to cope with geographically dispersed input, the convolutional neural network (CNN) [55] is the most efficient and frequently utilized.

In this study, we used a CNN-based framework that autonomously quantifies the feature map and classifies it. To the best of our knowledge, there is no literature on the use of CNN for the identification of floor mosaic patterns to date. Convolution, pooling and dense layers are three distinct categories of levels found in CNN. The convolution levels generate attributes from the incoming images by introducing certain specified filters. The generated feature vector is passed through a pooling layer to reduce the spatial size of the feature map. As a result, the network parameter count and computational cost are reduced. The dense level receives all the outputs from the preceding level and delivers one output to the following level from every neuron. The proposed CNN framework can be described as CPCCCPDD architecture, where C, P and D represent convolution, pooling and dense, respectively. The input image is fed to the first convolutional layer, which consists of 32 filters having size  $5 \times 5$ . This convolutional layer is followed by a max-pool layer with filter size  $3 \times 3$ . Then three convolutional layers having 16 filters of size  $3 \times 3$  each are fed in series. This is followed by another max-pool layer with filter size  $2 \times 2$ . There are two dense layers used in the proposed CNN framework: one is 45-dimensional dense and the second is 5-dimensional (output layer). The proposed CNN framework is depicted in Figure 2.



**Figure 2.** The proposed CNN architecture.

The number of pixels shifted across the incoming tensor is referred to as the stride. If the stride is set to 1, the filters/masks are moved one element at a time. If it is set to 2, then the mask will be shifted by two elements, and so on. Here, for both the convolution and pooling layers, the stride value of 1 is considered throughout the experiment. The dropout value of 0.5 was taken. The dropout helps to reduce the overfitting problem in the network. Before feeding to the dense layer, a batch normalization strategy is used to speed up the training process. The learning rate is taken as 0.001. The ‘Adam’ optimizer and ‘cross-entropy loss function’ are deployed in the proposed framework. In the convolutional

layers and the first dense layer, the rectified linear unit (ReLU) activation function is used, which can be formalized as:

$$f(n) = \max(0, n) \quad (1)$$

where  $n$  is the input to a neuron.

In the output layer, the activation function named ‘Softmax’ is used, which is provided in Equation (2).

$$\sigma(y)_i = \frac{e^{y_i}}{\sum_{l=1}^L e^{y_l}} \quad (2)$$

where  $y$  is the  $i$ th input vector of length  $l$ .

The number of parameters used in the CNN architecture is presented in Table 1. The total number of trainable parameters used is 617,491.

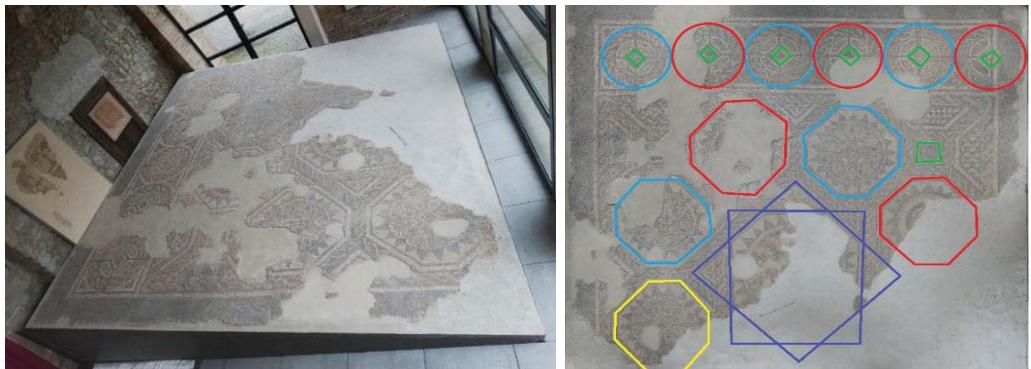
**Table 1.** Number of parameters used in the various levels of the CNN architecture for the presented design.

Layer	Dimension	#Parameters
Convolution 1	$196 \times 196 \times 32$	2432
Max-pool	$3 \times 3$	-
Convolution 2	$63 \times 63 \times 16$	4624
Convolution 3	$61 \times 61 \times 16$	2320
Convolution 4	$59 \times 59 \times 16$	2320
Max-pool	$2 \times 2$	-
Dense 1	45	605,565
Dense 2	5	230
Total		617,491

#### 4. Case Study

The deep learning (CNN) framework was applied and tested on a Roman mosaic discovered in Savignano sul Panaro, near the city of Modena (Italy), in 1897 during an archaeological excavation. This floor mosaic belongs to the ruins of a large late Roman building dated to the 5th century A.D. [56]. It originally measured about  $6.90 \text{ m} \times 4.50 \text{ m}$ , but less than half of its original surface is preserved. The Roman mosaic was removed for restoration and is now conserved in the birthplace house of the painter Giuseppe Graziosi (Savignano sul Panaro), who first documented its existence in 1897 (Figure 3, left).

The mosaic pattern is described in [57]. Its decorations present polychrome stone and terracotta tiles combined with emerald green and ruby red glass tiles. The mosaic shows a geometrical pattern of (originally) eight octagonal elements arranged around a larger central one, which consists of an eight-pointed star, formed by two superimposed squares to form a central octagon with irregular sides (in purple, in Figure 3, right). The central octagon has a circular motif with a white background containing a laurel wreath and, presumably, a figured center. The vertices of the star originate eight octagons, smaller in size, arranged in pairs of two on each side (in red, blue and yellow, in Figure 3, right), containing geometric and stylized plants that alternate with Solomon’s knots. The external octagons are only partially preserved, but all of them have internal circular motifs, with a border of pointed triangles in black on white. The space between the octagons and the side walls is filled with different polygonal and triangular forms. At the top, six circles (five full circles and one half-circle) alternate intertwined motifs with a red and black background, surrounding a central square.

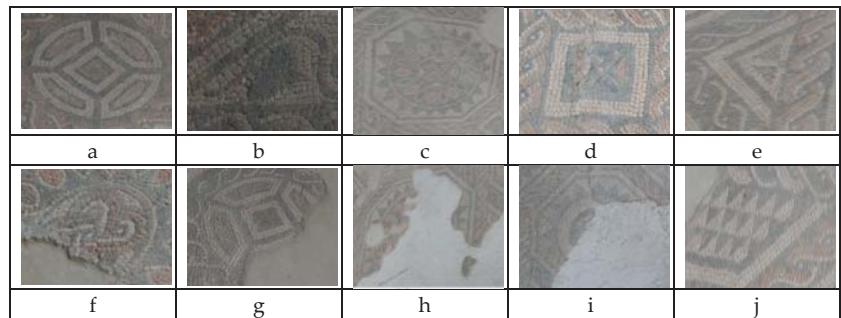


**Figure 3.** **Left:** The final location at the “Casa Natale Giuseppe Graziosi” in Savignano sul Panaro (Modena, Italy) (photo credits: Marianna Grandi, Italy). **Right:** Auxiliary geometric elements built on the orthophoto of the Roman mosaic, to highlight the geometric forms and their arrangement.

A close-range photogrammetric model of the Roman mosaic is developed by means of 115 photos (standard compact camera Nikon P310 (Nikon, Tokyo, Japan), 16.1MP CMOS sensor, sensor size: 1/2.3" (~6.16 mm × 4.62 mm), max. image resolution 4608 × 3456) thanks to Agisoft Metashape Professional (Version 1.6.3). In this software, the 3D model is also scaled to its natural size using as references the sides of the inclined support of the mosaic (see Figure 3, left), whose dimensions are known. The final model consists of a detailed textured 3D model of the mosaic, which shows the arrangements of the tiles, their edges and some planar issues due to its state of conservation, as well as the geometric forms and their arrangements.

The 3D model supported the generation of images showing the mosaic geometric forms in two ways: Firstly, from the 3D model, the Agisoft Metashape Pro software developed an orthophoto, which is a computer-generated image of the whole artifact that has been corrected for any geometric distortions. In particular, it is obtained as a parallel projection of the view of a photogrammetric textured model taken along a predetermined plane [58]. During the transformation from a 2D perspective view into an orthophoto, each photo is rectified (i.e., it is an orthogonal projection of the real photo on the mosaic plane); therefore, it is no longer deformed by perspective. Conversely, the “real” photo is influenced by perspective, as seen by the human eye. Therefore, we obtained a set of 115 photographic images corrected and rectified, from which we could extract the images of geometrical forms to be classified by the deep learning algorithm.

Secondly, from the 3D model, we extracted and isolated additional image samples depicting each of the geometric forms to be analyzed. By simply rotating, translating and zooming the 3D models, we obtained images of the same geometric form with multiple spatial orientations and, therefore, with multiple distortions. Some of these images are shown in Figure 4.



**Figure 4.** Some image samples of the mosaic forms: (a) circle, (b) leaf, (c) octagon, (d) square, (e) triangle; incomplete geometric forms from (f–j).

## 5. Experiments

### 5.1. Dataset

In this work, a dataset of images of the geometric forms of the floor mosaic was developed. Five different mosaic forms (i.e., tile patterns) were considered in this set: circles, triangles, leaves, octagons and squares.

The dataset contains 407 mosaic images, including 103 images of circles, 79 of octagons, 71 of squares, 137 of triangles and 17 of leaves. Figure 4 shows the mosaic image samples from the developed dataset, in which the mosaic tiles are arranged in patterns originating geometric forms. A circle-shaped motif of the mosaic texture is presented in Figure 4a. Similarly, (b) shows a leaf-shaped mosaic, (c) shows an octagon-shaped mosaic, (d) shows a square-shaped mosaic and (e) shows a triangle-shaped mosaic. The dataset contains images of different size such as  $540 \times 244$ ,  $352 \times 566$ ,  $737 \times 535$ ,  $869 \times 760$  and  $1535 \times 735$ . Since the image sizes were different, we normalized the height and width and set the size of  $200 \times 200$  before feeding to the deep learning-based framework. The images were captured in low lighting conditions. In addition, some of the images show forms that are not completely observable. In the second row (f–j) of Figure 4, the incomplete forms of the mosaic are shown. Some incomplete circular forms are shown as semicircles in (f) and (g), and inside the circle, there is a pattern of squares (g). The remaining parts of octagonal mosaic motifs are shown in (h) and (i). In (j), there are many triangle-shaped motifs within a large square, whose actual patterns are difficult to identify. The correct identification of the mosaic forms in the dataset is complicated as the data suffer from incomplete structure, poor light condition, blurriness and low volume of data.

### 5.2. Evaluation Protocol

We used an n-fold cross-validation technique to test the efficiency of our system. In this cross-validation approach, the entire dataset was divided into n parts: training set and test set. The test set is considered as one of the  $n$  parts, whereas the rest ( $n - 1$ ) are considered as the training set. In the next iteration, out of ( $n - 1$ ) sets, one of the sets is considered as a test set (different from before), and the remaining ( $n - 1$ ) parts are considered as the training set, and so on. This process is repeated n times. Various metrics such as accuracy, precision, recall and F-score, used to assess the effectiveness of the system, are computed as:

$$\text{Accuracy} = ((tp + tn) / (tp + fp + fn + tn)) \quad (3)$$

$$\text{Precision} = tp / (tp + fp) \quad (4)$$

$$\text{Recall} = tp / (tp + fn) \quad (5)$$

$$\text{F-score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (6)$$

where the true positive, false positive, false negative and true negative parameters are represented by  $tp$ ,  $fp$ ,  $fn$  and  $tn$ , respectively.

### 5.3. Results and Analysis

Table 2 presents the performance metrics obtained with a batch size equal to 100 and for 100 epochs. It shows that the highest accuracy of 93.61% was obtained for the 10-fold cross-validation. If the number of folds increases, the accuracy decreases.

**Table 2.** The performance of the CNN architecture in different folds of cross-validation with a batch size equal to 100 and for 100 epochs.

#Fold	Accuracy	Precision	Recall	F-Score
5	91.40	0.9348	0.8912	0.9100
7	89.68	0.9108	0.8475	0.8714
10	93.61	0.9529	0.9236	0.9367
12	89.19	0.9159	0.8879	0.8960

With the 10-fold cross-validation and the batch size equal to 100, the performance of the system was analyzed by changing the number of epochs. Table 3 shows the results of the performance considering from 200 to 700 epochs with intervals of 100 epochs. It shows that, at 500 epochs, the highest values of accuracy (97.05%), recall (0.9658) and F-score (0.9651) were obtained.

**Table 3.** The performance evaluation by changing the number of epochs with the 10-fold cross-validation and a batch size equal to 100.

Epoch	Accuracy	Precision	Recall	F-Score
200	0.941	0.9563	0.9252	0.9395
300	96.81	0.9742	0.9599	0.9667
400	95.82	0.9658	0.9505	0.9578
500	97.05	0.9645	0.9658	0.9651
600	93.37	0.9459	0.9189	0.9313
700	91.89	0.947	0.9067	0.9246

Further experimentation was carried out by increasing the batch size from 50 to 250 with 50 batch intervals, keeping the 10-fold cross-validation and 500 epochs. The performance metrics are presented in Table 4, which shows that increasing the batch size did not improve the performance. The same accuracy was obtained for the batch sizes equal to 50 and 100, but higher precision and F-score were found for the batch size equal to 50.

**Table 4.** For 500 epochs and 10-fold cross-validation, the metrics were calculated by increasing the batch size from 50 to 250.

Batch	Accuracy	Precision	Recall	F-Score
50	97.05	0.9760	0.9632	0.9693
100	97.05	0.9645	0.9658	0.9651
150	93.61	0.9472	0.9253	0.9354
200	90.37	0.9021	0.8913	0.8966
250	92.87	0.9438	0.9002	0.9184

The confusion matrix (in Table 5) was explored for the 10-fold cross-validation, a batch size equal to 50 and 500 epochs.

**Table 5.** The confusion matrix of the accuracy corresponding to the five floor mosaic patterns.

	Circles	Leaves	Octagons	Squares	Triangles
Circles	97.08	0	0	0.019	0.009
Leaves	0	94.11	0	0	0.058
Octagons	0.012	0	97.46	0	0.012
Squares	0.056	0	0	94.33	0
Triangles	0.007	0	0.007	0	98.54

The confusion matrix shows that the triangle patterns present the highest accuracy (98.54%), followed by the octagons (97.46%), the circles (97.08%), the squares (94.36%) and the leaves (94.11%). The errors in identification were generated because of poor illumination, noise, blurriness and improper/incomplete geometry of the floor mosaic patterns.

#### 5.4. Comparison

The performance of the system was compared to standard CNN architectures. Here, four different architectures were considered, namely VGG19 [59], MobileNetV2 [60], ResNet50 [61] and InceptionV3 [62]. VGG19, MobileNetV2, ResNet50 and InceptionV3 networks are 19, 53, 50 and 48 layers deep, while the proposed network consists of only nine layers. Instead of applying deep networks, the proposed framework gives us better performance. The comparison results are shown in Table 6.

**Table 6.** Comparison of the proposed framework with standard CNN-based networks.

Network	Accuracy (%)	Precision	Recall	F-Score
VGG19	93.90	0.9409	0.9278	0.9343
MobileNetV2	89.78	0.9056	0.8860	0.8956
ResNet50	84.67	0.8478	0.8408	0.8442
InceptionV3	78.55	0.7720	0.7803	0.7761
<b>Proposed</b>	<b>97.05</b>	<b>0.9645</b>	<b>0.9658</b>	<b>0.9651</b>

#### 6. Discussion and Conclusions

This paper presents a framework for geometric form analysis based on images extracted from a close-range photogrammetric model of an artifact (floor mosaic) and deep learning (CNN) algorithm. From the digital model of the mosaic, an orthophoto was obtained, which the photogrammetric software generated by rectifying the photos used in photogrammetry. Therefore, two sets of photos were collected in a dataset: the original photos, affected by perspective, useful for obtaining images of the deformed geometric forms of the mosaic and, on the other hand, the rectified version of the same photos with the geometric forms projected on the floor plane and so not deformed. Moreover, additional images can be obtained by simply rotating, translating and zooming the 3D model of the mosaic, generating other images with geometric forms differently deformed.

The deep learning algorithm analyzed the entire dataset consisting of 407 (normalized) images, in particular, 103 images of circles, 79 images of octagons, 71 images of squares, 137 images of triangles and 17 images of leaves. The geometric forms in the mosaic are made by arrangements of tiles, which caused jagged contours and irregularities in the geometric forms to be analyzed; moreover, there were cracks and improper/incomplete geometry of the mosaic elements, which were sometimes due to unevenness in the ground

or the elements having been destroyed in the past. Moreover, some of the photos showing the mosaic forms present noise and blurs, sometimes due to poor illumination.

Despite all these defects, the algorithm is able to identify and classify more than 94% of the forms in each category, and the method has proved to be robust enough to analyze the mosaic geometric forms chosen as a case study. Furthermore, the performance of the proposed method was compared with standard deep architectures that deployed a larger number of convolutions and pooling layers than the proposed method. Instead, we achieved good accuracy using the proposed lightweight architecture.

Concerning the selected case study, the proposed method has proved to be capable of extracting and classifying data from this kind of artwork. The dataset consists of various images related to five geometric forms that are repeated in the mosaic using different arrangements of tiles, colors and orientation, usually incomplete or separated by diameters, diagonals or simply by including smaller geometric forms in larger ones. Despite all these differences among the same kinds of geometric forms, the CNN architecture has proven to be capable of classifying the five geometric forms with high accuracy; therefore, we confidently believe that it can be easily generalized to other mosaics with similar forms and patterns. As it was not possible to test it as part of this research activity, testing the CNN algorithm with other mosaics will be planned as future work.

Additional future works will consist in the analysis of mosaics and other artworks that are not flat but 3D-shaped in space, such as curved walls, domes and vaults. In addition, the method can originate a software tool for processing and analyzing fine arts data in a more automated way.

**Author Contributions:** Conceptualization, F.G., S.M.O. and M.Z.; methodology, F.G., S.M.O. and M.Z.; software, M.G. and S.M.O.; validation, M.G. and S.M.O.; data curation, F.G., M.G. and S.M.O.; writing—original draft preparation F.G., S.M.O. and M.Z.; writing—review and editing, F.G., S.M.O. and M.Z.; supervision, F.G., S.M.O. and M.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** Mária Ždímalová acknowledges the financial support of the Slovak Scientific Grant VEGA No. 1/0006/19.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Barni, M.; Pelagotti, A.; Piva, A. Image processing for the Analyses and Conservation of Paintings: Opportunity and challenges. *IEEE Signal Process. Mag.* **2005**, *22*, 141–144. [[CrossRef](#)]
- Cornelis, B. Image processing for art Investigation. *Electron. Lett. Comput. Image Anal.* **2014**, *14*, 1–7. [[CrossRef](#)]
- Johnson, C.R., Jr.; Hendriks, E.J.; Berezhnyi, I.; Brevdo, E.; Huges, S.M.; Daubechies, I.; Li, J.; Postma, E.; Wang, J.Z. Image processing for artist identification, Computerizes Analysis of Vincent van Gogh. *IEEE Signal Process. Mag.* **2008**, *25*, 37–48. [[CrossRef](#)]
- Bartolini, F.; Cappellini, V.; Del Mastio, A.; Piva, A. Applications of image processing technologies to fine arts. *Opt. Metrol. Arts Multimed.* **2003**, *5146*. [[CrossRef](#)]
- Berezhnoy, I.E.; Postma, E.O.; van den Herik, H.J. Computerized visual analysis of paintings. *Proc. Int. Conf. Assoc. Hist. Comput.* **2005**, 28–32. Available online: <https://repository.ubn.ru.nl/bitstream/handle/2066/32358/32358.pdf?sequence=1#page=29> (accessed on 14 August 2021).
- Teixiera, G.N.; Feitosa, R.Q.; Paciornik, S. *Pattern Recognition Applied in Fine Art Authentication*; Catholic University of Rio de Janeiro: Rio de Janeiro, Brazil, 2002; Available online: [http://www.lvc.ele.puc-rio.br/users/raul\\_feitosa/publications/2002/Pattern%20recognition%20applied.pdf](http://www.lvc.ele.puc-rio.br/users/raul_feitosa/publications/2002/Pattern%20recognition%20applied.pdf) (accessed on 14 August 2021).
- Amura, A.; Aldini, A.; Pagnotta, S.; Salerno, E.; Tonazzini, A.; Triolo, P. Analysis of Diagnostic Images of Artworks and Feature Extraction: Design of a Methodology. *J. Imaging* **2021**, *7*, 53. [[CrossRef](#)]
- Daffara, C.; Ambrosini, D.; Di Biase, R.; Fontana, R.; Paoletti, D.; Pezzati, L.; Rossi, S. Imaging data integration for painting diagnostics. In Proceedings of the O3A: Optics for Arts, Architecture, and Archaeology II, Munich, Germany, 17–18 June 2009; Volume 7391. [[CrossRef](#)]
- Cappellini, V.; Barni, M.; Corsini, M.; Rosa, A.D.; Piva, A. Artshop: An art-oriented image processing tool for cultural heritage applications. *J. Visual. Comput. Animat.* **2003**, *14*, 149–158. [[CrossRef](#)]
- Milidiu, R.; Renteria, R. *Projeto Pinçelada*; Pontifícia Universidade Católica do Rio de Janeiro: Rio de Janeiro, Brazil, 1998.

11. Pei, S.-C.; Zeng, Y.-C.; Chang, C.-H. *Virtual Restoration of Ancient Chinese Paintings Using Color contrast Enhancement and Lacuna Texture Synthesis*; IEEE: Manhattan, NY, USA, 2004; Volume 13, pp. 416–429.
12. Bellavia, F.V.; Colombo, C. Color correction for image stitching by monotone cubic spline interpolation. In Proceedings of the 7th Iberian Conference on Pattern Recognition and Image Analysis, Santiago de Compostela, Spain, 17–19 June 2015; Volume 9117, pp. 165–172. [[CrossRef](#)]
13. Zhang, D.; Islam, M.; Lu, G. A review on automatic image annotation techniques. *Pattern Recognit.* **2012**, *45*, 346–362. [[CrossRef](#)]
14. Cornelis, B.; Dooms, A.; Cornelis, J.; Schelkens, P. Digital canvas removal in paintings. *Signal Process.* **2012**, *92*, 1166–1171. [[CrossRef](#)]
15. Yin, R.; Dunson, D.; Cornelis, B.; Brown, B.; Ocon, N.; Daubechies, I. Digital Cradle Removal in X-ray Images of Art Paintings. In Proceedings of the IEEE International Conference on Image Processing, Paris, France, 27–30 October 2014.
16. Cornelis, B.; Ruzic, T.; Gezels, E.; Dooms, A.; Pizurica, A.; Platisa, L.; Cornelis, J.; Martens, M.; De Mey, M.; Daubechies, I. Crack detection and in painting for virtual restoration of paintings: The case of the Ghent Altarpiece. *Signal Process.* **2013**, *93*, 605–619. [[CrossRef](#)]
17. Cornelis, B.; Yang, Y.; Vogelstein, J.T.; Dooms, A.; Daubechies, I.; Dunson, D. Bayesian crack detection in ultra high resolution multimodal images of paintings. In Proceedings of the 18th International Conference on Digital Signal Processing (DSP), Santorini, Greece, 1–3 July 2013; pp. 1–8. [[CrossRef](#)]
18. Cornelis, B.; Dooms, A.; Munteanu, A.; Cornelis, J.; Schelkens, P. Experimental study of canvas characterization for paintings. In *Computer Vision and Image Analysis of Art*; SPIE Press: San Jose, CA, USA, 2010.
19. Barni, M.; Cappellini, V.; Mecocci, A. The use of different metrics in vector median filtering: Application to fine arts and paintings. In Proceedings of the 6th European Signal Processing Conference, Brussels, Belgium, 25–28 August 1992; pp. 1485–1488.
20. Lu, C.S.; Chung, P.C.; Chen, C.F. Unsupervised texture segmentation via wavelet transformation. *Pattern Recognit.* **1997**, *30*, 729–742. [[CrossRef](#)]
21. Chen, C.C.; Chen, C.C. Filtering methods for texture discrimination. *Pattern Recognit. Lett.* **1999**, *20*, 783–790. [[CrossRef](#)]
22. Castellano, G.; Vessio, G. Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. *Neural Comput. Appl.* **2021**, *1*–20. [[CrossRef](#)]
23. Castellano, G.; Vessio, G. Deep convolutional embedding for digitized painting clustering. In Proceedings of the International Conference on Pattern Recognition, Virtual, Milan, 10–15 January 2021. [[CrossRef](#)]
24. Stork, D.G. From Digital Imaging to Computer Image Analysis of Fine Art. In *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*; Huang, F., Wang, R.C., Eds.; Springer: Berlin/Heidelberg, Germany, 2010. [[CrossRef](#)]
25. Basvaprasad, B.; Hegadi, R.S. A survey on traditional and graph theoretical technique for image segmentation. *Inter. J. Comput. Appl.* **2014**, *957*, 8887.
26. Bazi, Y.; Bruzzone, L.; Melgani, F. Image thresholding based on the EM algorithm and the generalized Gaussian distribution. *Pattern Recognit.* **2007**, *40*, 619–634. [[CrossRef](#)]
27. Davies, E.R. Chapter 4—Thresholding Techniques. In *Computer and Machine Vision*, 4th ed.; Davies, E.R., Ed.; Academic Press: Cambridge, MA, USA, 2012; pp. 82–110. [[CrossRef](#)]
28. Callara, A.L.; Magliaro, C.; Ahluwalia, A.; Vanello. A Smart Region-Growing Algorithm for Single-Neuron Segmentation From Confocal and 2-Photon Datasets. *Front. Neuroinform.* **2020**, *14*, 8–12. [[CrossRef](#)]
29. Maeda, J.; Ishikawa, C.; Novianto, S.; Tadehara, N.; Suzuki, Y. Rough and accurate segmentation of natural color images using fuzzy region-growing algorithm. In Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, Spain, 3–7 September 2000.
30. Peng, B.; Zhang, L.; Zhang, D. A survey of graph theoretical approaches to image segmentation. *Pattern Recognit.* **2013**, *46*, 1020–1038. [[CrossRef](#)]
31. Magzhan, K.; Matjani, H. A review and evaluations of shortest path algorithm. *Int. J. Sci. Technol. Res.* **2013**, *2*, 99–104.
32. Yi, F.; Moon, I. Image segmentation: A survey of graph-cut methods. In Proceedings of the IEEE International Conference on Systems and Informatics, Yantai, China, 19–20 May 2012.
33. Han, P.; Li, Z.; Gong, J. Effects of Aggregation Methods on Image Classification. In *Technology for Earth Obs. Geospatial*; Li, D., Shan, J., Gong, J., Eds.; Springer: Boston, MA, USA, 2010; pp. 271–288. [[CrossRef](#)]
34. Guedj, B.; Rengot, J. Non-linear Aggregation of Filters to Improve Image Denoising. In *Advances in Intelligent Systems and Computing*; Arai, K., Kapoor, S., Bhatia, R., Eds.; Springer: Cham, Switzerland, 2020; Volume 1229. [[CrossRef](#)]
35. Hao, X.; Zhang, G.; Ma, S. Deep learning. *Int. J. Semant. Comput.* **2020**, *10*, 417–439. [[CrossRef](#)]
36. Ghosh, M.; Mukherjee, H.; Obaidullah, S.M.; Santosh, K.C.; Das, N.; Roy, K. Identifying the presence of graphical texts in scene images using CNN. In Proceedings of the 2019 International Conference on Document Analysis and Recognition, Sydney, Australia, 20–25 September 2019.
37. Castellano, G.; Vessio, G. A Brief Overview of Deep Learning Approaches to Pattern Extraction and Recognition in Paintings and Drawings. In Proceedings of the 25th International Conference on Pattern Recognition Workshops, Milan, Italy, 10–11 January 2021. [[CrossRef](#)]
38. Castellano, G.; Lella, E.; Vessio, G. Visual link retrieval and knowledge discovery in painting datasets. *Multimed. Tools Appl.* **2021**, *80*, 6599–6616. [[CrossRef](#)]

39. Sharma, D.; Gupta, N.; Chattopadhyay, C.; Mehta, S. Daniel: A deep architecture for automatic analysis and retrieval of building floor plans. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; pp. 420–425.
40. Osuna-Coutiño, J.D.J.; Martinez-Carranza, J. Structure extraction in urbanized aerial images from a single view using a CNN-based approach. *Int. J. Remote Sens.* **2020**, *41*, 8256–8280. [CrossRef]
41. Ziran, Z.; Marinai, S. Object detection in floor plan images. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*; Springer: Cham, Switzerland, 2018; pp. 383–394.
42. Gómez-Ríos, A.; Tabik, S.; Luengo, J.; Shihavuddin, A.S.M.; Krawczyk, B.; Herrera, F. Towards highly accurate coral texture images classification using deep convolutional neural networks and data augmentation. *Expert Syst. Appl.* **2019**, *118*, 315–328. [CrossRef]
43. Feng, S.; Chen, Q.; Gu, G.; Tao, T.; Zhang, L.; Hu, Y.; Wei, Y.; Zuo, C. Fringe pattern analysis using deep learning. *Adv. Photonics* **2019**, *1*, 025001. [CrossRef]
44. Sandelin, F. Semantic and Instance Segmentation of Room Features in Floor Plans Using Mask R-CNN. Master’s Thesis, Uppsala Universitet, Uppsala, Sweden, 2019. Available online: <http://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A1352780&ddswid=8811> (accessed on 14 August 2021).
45. Vilnrotter, F.M.; Nevatia, R.; Price, K.E. Structural analysis of natural textures. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*; IEEE: New York, NY, USA, 1986; Volume 1, pp. 76–89.
46. Adami, A.; Fassi, F.; Fregonese, L.; Piana, M. Image-Based Techniques For the Survey of Mosaics in the St Mark’s Basilica in Venice. *Virtual Archaeol. Rev.* **2018**, *9*, 1–20. [CrossRef]
47. Doria, E.; Picchio, F. Techniques For Mosaics Documentation Through Photogrammetry Data Acquisition. The Byzantine Mosaics Of The Nativity Church. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *5*, 2.
48. Fioretti, G.; Acquafrredda, P.; Calò, S.; Cinelli, M.; Germanò, G.; Laera, A.; Moccia, A. Study and Conservation of the St. Nicola’s Basilica Mosaics (Bari, Italy) by Photogrammetric Survey: Mapping of Polychrome Marbles, Decorative Patterns and Past Restorations. *Stud. Conserv.* **2020**, *65*, 160–171. [CrossRef]
49. Fazio, L.; Lo Brutto, M.; Dardanelli, G. Survey and virtual reconstruction of ancient roman floors in an archaeological context. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *42*, 511–518. [CrossRef]
50. Zitova, B.; Flusser, J.; Ľroubek, F. An application of image processing in the medieval mosaic conservation. *Pattern Anal. Appl.* **2004**, *7*, 18–25. [CrossRef]
51. Felicetti, A.; Paolanti, M.; Zingaretti, P.; Pierdicca, R.; Malinverni, E.S. Mo.Se.: Mosaic image segmentation based on deep cascading learning. *Virtual Archaeol. Rev.* **2021**, *12*. [CrossRef]
52. Benyoussef, L.; Derrode, S. Analysis of ancient mosaic images for dedicated applications. In *Digital Imaging for Cultural Heritage Preservation—Analysis, Restoration, and Reconstruction of Ancient Artworks*; Filippo., S., Sebastian, B., Giovanni, G., Eds.; CRC Press: Boca Raton, FL, USA, 2017; 523p.
53. Falomir, Z.; Museros, L.; Gonzalez-Abril, L.; Velasco, F. Measures of similarity between qualitative descriptions of shape, colour and size applied to mosaic assembling. *J. Vis. Communun. Image Represent.* **2013**, *24*, 388–396. [CrossRef]
54. Ghosh, M.; Mukherjee, H.; Obaidullah, S.M.; Santosh, K.C.; Das, N.; Roy, K. LWSINet: A deep learning-based approach towards video script identification. *Multimed. Tools Appl.* **2021**, *1*–34. [CrossRef]
55. Ghosh, M.; Roy, S.S.; Mukherjee, H.; Obaidullah, S.M.; Santosh, K.C.; Roy, K. Understanding movie poster: Transfer-deep learning approach for graphic-rich text recognition. *Vis. Comput.* **2021**, *1*–20. [CrossRef]
56. Gherardini, F.; Santachiara, M.; Leali, F. Enhancing heritage fruition through 3D virtual models and augmented reality: An application to Roman artefacts. *Virtual Archaeol. Rev.* **2019**, *10*, 67–79. [CrossRef]
57. Santachiara, M.; Gherardini, F.; Leali, F. An Augmented Reality Application for the Visualization and the Pattern Analysis of a Roman Mosaic. In *IOP Conference Series: Materials Science and Engineering*. Kuala Lumpur, Malaysia, 13–14 August 2018; IOP Publishing: Bristol, UK, 2018; Volume 364, p. 012094.
58. Ippolito, A.; Cigola, M. *Handbook of Research on Emerging Technologies for Digital Preservation and Information Modeling*; Information Science Reference: Hersey, PA, USA, 2017.
59. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
60. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
61. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
62. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.



Article

# Multimodal Emotion Recognition from Art Using Sequential Co-Attention

Tsegaye Misikir Tashu <sup>1,2,\*</sup>, Sakina Hajiyeva <sup>1</sup> and Tomas Horvath <sup>1,3</sup>

<sup>1</sup> Department of Data Science and Engineering (T-Labs), Faculty of Informatics, Eötvös Loránd University, Pázmány Péter Sétány 1/C, 1117 Budapest, Hungary; Hajiyeva44@gmail.com (S.H.); tomas.horvath@inf.elte.hu (T.H.)

<sup>2</sup> College of Informatics, Kombolcha Institute of Technology, Wollo University, Kombolcha 208, Ethiopia

<sup>3</sup> Faculty of Science Institute of Computer Science, Pavol Jozef Šafárik University, Jesenná 5, 040 01 Košice, Slovakia

\* Correspondence: misikir@inf.elte.hu

**Abstract:** In this study, we present a multimodal emotion recognition architecture that uses both feature-level attention (sequential co-attention) and modality attention (weighted modality fusion) to classify emotion in art. The proposed architecture helps the model to focus on learning informative and refined representations for both feature extraction and modality fusion. The resulting system can be used to categorize artworks according to the emotions they evoke; recommend paintings that accentuate or balance a particular mood; search for paintings of a particular style or genre that represents custom content in a custom state of impact. Experimental results on the WikiArt emotion dataset showed the efficiency of the approach proposed and the usefulness of three modalities in emotion recognition.



**Citation:** Tashu, T.M.; Hajiyeva, S.; Horvath, T. Multimodal Emotion Recognition from Art Using Sequential Co-Attention. *J. Imaging* **2021**, *7*, 157. <https://doi.org/10.3390/jimaging7080157>

Academic Editors: Gennaro Vessio, Giovanna Castellano and Fabio Bellavia

Received: 23 July 2021

Accepted: 17 August 2021

Published: 21 August 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Art is an imaginative human creation that should be appreciated, make people think, and evoke an emotional response [1–3]. Emotion is a psycho-physiological process that can be triggered by conscious and/or unconscious perceptions of objects and situations and is related to a variety of factors such as mood, temperament, personality, disposition, and motivation [2,3]. Emotions are very important in human decision making, interaction and cognitive processes [4]. As technology advances and our understanding of emotions grows, so does the need for automatic emotion recognition systems [2]. Automatic emotion recognition has been used for various applications including human–computer interactions [5], surveillance [6], robotics, gaming, entertainment, and more.

Initial work on emotion recognition was mostly carried out using unimodal [7,8] approaches. Unimodal modality can correspond to facial expressions, voice, text, posture, gaits, or physiological signals. This was followed by multimodal emotion recognition [9,10], where different modalities were used and combined in various ways to derive emotions.

However, most of the work on automatic analysis of artworks has focused on inferring painting styles [11], investigating influences between artists and art movements [12], distinguishing authentic drawings from imitations, automatically generating artworks [13], and evaluating evoked emotions [14,15]. There are also attempts to develop approaches to analyze people's emotional experiences in response to artworks [14,15]. Most of these studies use computer vision and machine learning approaches to emotionally categorize artworks [14,15] and identify the parts of paintings that are responsible for evoking certain emotions [16].

Automatic detection of emotions evoked by art paintings is of significant importance as the results can be used to group art paintings according to the emotions they evoke,

to provide painting recommendations that accentuate or balance a particular mood, and to find art paintings of a particular style or genre that represent user-defined content in a user-defined state of effect [1–3].

We proposed a co-attention-based multimodal emotion recognition model that jointly identifies reasons from all modalities used and a weighted modality fusion that provides feature-level system fusion and applies weighted modality scores over the extracted features to indicate the importance of the different modalities. We compared our approach to several baseline methods by testing the performance on the WikiArt emotion dataset [1], a benchmark dataset for emotion recognition in art. Our models can be used if the two modalities, namely the image (painting) and title (textual description), are provided. The third modality which is the emotion category is not possible to collect every time the model is used as its values come from the expert judgments. As the model was trained using the three modalities and to avoid any bugs during the deployment due to the missing category modality, we included a function that initializes the category modality into some value drawn randomly from a uniform distribution when the category modality is not present. The contribution of this paper can be summarized as follows:

1. We proposed a co-attention-based multimodal emotion recognition approach that aims to use information from the painting, title, and emotion category channels via weighted fusion to achieve more robust and accurate recognition;
2. An experiment was carried on the dataset collected and provided for emotion recognition, which is publicly available;
3. The proposed approach result was compared with the latest state-of-the-art approaches and also with other baseline approaches based on deep learning methods.

The rest of the paper is organized into five sections. Section 2 describes related works that are relevant to our research. Section 3 presents the proposed sequential multimodal fusion model architecture and Section 4 presents the overall experimental settings, implementation, and evaluation of the proposed system and results. Finally, Section 5 presents the conclusion.

## 2. Related Work

Emotion detection and sentiment analysis has been an area of interest for many decades and has always attracted attention in multiple fields using computer vision and natural language processing techniques. Depending on the number (uni- and multimodal) and types of modalities (speech, text, video, image), there have been some major improvements in the topic of emotion detection and sentiment analysis. In this section, we will focus on the most recent findings for unimodal and multimodal emotion recognition by discussing recent developments in techniques and approaches for each modality type.

### 2.1. Unimodal Approaches

The first attempts to identify human emotions were mostly unimodal. The most commonly studied modalities are facial expressions [7], speech or vocal expressions [17], body gestures [18], and physiological signals such as respiratory and cardiac signals [8]. Recent work in the field of unimodal emotion recognition agrees that building a model that can better capture the context and sequential nature of the input can significantly improve performance in the difficult task of emotion recognition. It has been shown that using a recurrent neural network-based classifier that can learn to create a more informative latent representation of the target as a whole significantly improves final performance. Based on this assumption, a deep recurrent neural network architecture was proposed to detect discrete emotions in a tweet dataset [19]. An interaction-aware attention network (IAAN) that incorporates contextual information into the learned voice representation through an attentional mechanism was proposed by Sung-Lin et al. [20]. The performance shows significant improvement over previously shown state-of-the-art and baseline methods and provides one of the best emotion recognition results [20].

## 2.2. Multimodal Approaches

As human beings, we usually rely on multiple factors such as intonation (speech), facial expression (visual modality), and contextual meaning of words (text) to detect emotions. For this reason, it is undeniably naive to expect unimodal models to outperform humans in emotion recognition and sentiment analysis. To be truly successful in emotion recognition, it is important to consider all possible mixtures of modalities. Multimodal emotion recognition is a field with many ideas and approaches and, in this part, we will focus on blending the modalities of speech, text and video. Multimodal emotion recognition has been studied using classifiers such as Support Vector Machines (SVMs) and linear and logistic regressions [21,22]. With the development of larger datasets, deep learning architectures have been developed and explored [23–26].

Shenoy and Sardana proposed context-aware emotion recognition that captures context across all modalities, bridging the gap in using the context of different inputs by using a recurrent neural network [9]. Although fusion mechanism is a popular approach in multimodal analysis, there are still some exceptions in using fusion. Features from different modalities were trained individually based on multiple classifiers. Emotion features are fused using beam search fusion learning from the beam search method [27]. In one of the recent works, instead of independently fusing the knowledge from different modalities, the attention mechanism was introduced to combine the information to perform emotion classification [10].

Pan, Zexu et al. [28] proposed a multimodal attention network (MMAN) that makes use of visual and textual signals in speech emotion recognition. Their experiment showed that identifying speech emotions profits immensely from visual and textual signals.

Siriwardhana et al. [29] used the pre-trained “BERT-like” architecture for self-supervised learning (SSL) to represent language and text modalities to learn language emotions. Their method showed that a shallow-fusion simplifies the overall structure and strengthens complex fusion mechanisms. Liu, Gaojun et al. [30] introduced a multimodal music emotion grouping approach based on music audio and lyrics. They used the LSTM network for audio modality and Bert for lyrics to describe the emotions of lyrics, which essentially addresses long-term dependency. The neural network is implemented based on linear weighted decision-making stage fusion, which increases efficiency.

## 2.3. Emotion Recognition from Art

Yanulevskaya et al. [16] proposed an approach to categorize emotions from art paintings based on an aggregation of local image statistics and SVM. Machajdik et al. [31] presented a unified framework for classifying artworks by combining low-level visual features with high-level concepts from psychology and art theory. The paper by Yanulevskaya et al. [32] introduced a “bag-of-visual-words” model combined with SVM to classify abstract paintings into positive or negative emotions. Sartori et al. [33] introduced a general learning method for emotion recognition in abstract paintings that integrates both visual and textual information.

For various reasons, most work on emotion recognition in art paintings is unimodal. Using information from different modalities could increase the model accuracy in emotion recognition. In this work, we propose a co-attention-based multimodal emotion recognition approach that aims to use information from the painting, title, and emotion category channels via weighted fusion to achieve more robust and accurate recognition.

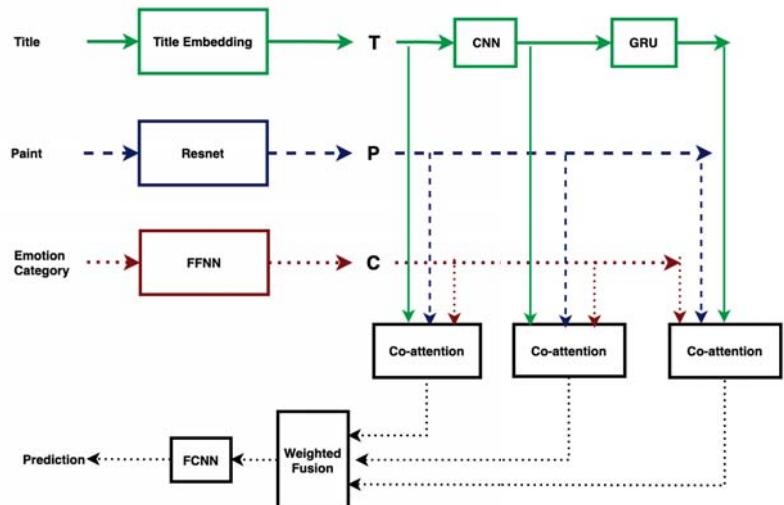
## 3. The Proposed Sequential Multimodal Fusion Model

Figure 1 shows the architecture of our sequential attention-based multimodal model with weighted fusion approach. Here, the title of the paint, the paint (image) and the emotion category attributes are treated as the three modalities. The weighted modality fusion technique is used to fully utilize the three modalities, and it has been shown that the model performance can be enhanced by adding the high-level concept [3,34]. In the

following, the text vector, the image vector and the emotion category vector are defined and the weighted fusion technique is briefly introduced.

For the imaging modality, the pre-trained and fine-tuned ResNet [35] model is used to obtain  $14 \times 14$  regional vectors of the art image, defined as the raw image vectors averaged to obtain the image vector. A Convolutional Neural Network (CNN) and a Bi-directional Gated Recurrent Unit (Bi-GRU) is used to obtain the text vectors. The word-level and n-gram level text vectors are processed using Bi-GRU to obtain the title level text feature vector. We used a three-layer feedforward neural network to obtain the emotion category feature vectors from the emotion category attributes.

To use multimodal information from all modalities and to refine the representation of all modalities, we proposed to use a sequential-based attention layer [36,37] that learns a new refined weighted representation for each of the input modalities. The refined vectors of the three modalities are combined in the modality fusion process to form a vector with weighted modality fusion [2,34] instead of simple concatenation. Finally, the fused vector is transferred to a three-layer fully connected neural network to obtain a classification result. The whole framework is shown in Figures 1 and 3. More details about our model can be found below.



**Figure 1.** The proposed attention-based model.

### 3.1. Image Feature Representation

The ResNet-50 [35] model is used to obtain representations of Art images. The last fully connected (FC) layer of the pre-trained model is chopped and replaced with a new one for the sake of model fine-tuning. Following the work of [34,36], an input image  $I$  is re-sized to  $448 \times 448$  and divided into  $14 \times 14$  regions. Each region  $I_i$  ( $i = 1, 2, \dots, 196$ ) is then sent through the ResNet model to obtain a regional feature representation, a.k.a., a raw image vector. The final image feature vector ( $P$ ) is obtained by the average of all regional image vectors.

$$P = \frac{\sum_{i=1}^{N_r} \text{ResNet}(I_i)}{N_r} \quad (1)$$

where  $\text{ResNet}(I_i)$  is the row image vector extracted via ResNet,  $N_r$  (set to 196 in this work) is the number of regions as in [36].  $P$  is the average of all regional image vectors.

### 3.2. Text Feature Representation

The sequence of word embeddings learned from the embedding layer was passed to a 1D convolution neural network for feature extraction at different levels. The resulting feature vector was further used to be fed to a Bi-GRU network layer to learn title level feature representation. GRU was recently introduced as an alternative to the long-short term memory (LSTM) model to make each recurrent unit to adaptively capture dependencies of different time scales [38,39]. Similarly to the LSTM unit, GRU has gating units that modulate the flow of information inside the unit, but without having a separate memory cell. The updates performed at each time step  $t \in \{1, \dots, T\}$  in a GRU are as follows:

Forward updates:

$$\vec{Z}_t = \text{sigmoid}(\vec{W}_z X_t + \vec{U}_z h_{t-1}) \quad (2)$$

$$\vec{r}_t = \text{sigmoid}(\vec{W}_r X_t + \vec{U}_r h_{t-1}) \quad (3)$$

$$\vec{h}_t = \tanh(\vec{W}_h X_t + \vec{U}_h(r_t \odot h_{t-1})) \quad (4)$$

$$\vec{h}_t = (1 - \vec{Z}_t) \odot \vec{h}_{t-1} + \vec{Z}_t \odot \vec{h}_t \quad (5)$$

Backward updates:

$$\overleftarrow{Z}_t = \text{sigmoid}(\overleftarrow{W}_z X_t + \overleftarrow{U}_z h_{t-1}) \quad (6)$$

$$\overleftarrow{r}_t = \text{sigmoid}(\overleftarrow{W}_r X_t + \overleftarrow{U}_r h_{t-1}) \quad (7)$$

$$\overleftarrow{h}_t = \tanh(\overleftarrow{W}_h X_t + \overleftarrow{U}_h(r_t \odot h_{t-1})) \quad (8)$$

$$\overleftarrow{h}_t = (1 - \overleftarrow{Z}_t) \odot \overleftarrow{h}_{t-1} + \overleftarrow{Z}_t \odot \overleftarrow{h}_t \quad (9)$$

where  $W_z, W_r, W_h, U_r, U_z, U_h, U_o$  are the weight matrices,  $\odot$  is an element-wise multiplication. The activation  $h_t$  at time  $t$  is a linear interpolation between the previous activation  $h_{t-1}$  and the candidate activation  $\hat{h}_t$ . An update gate  $Z_t$  decides how much the unit updates its activation or content. The reset gate ( $r_t$ ) is used to control access to the previous state  $h_{t-1}$  and compute a proposed update  $\hat{h}_t$ . When off ( $r_t$  close to 0), the reset gate effectively makes the unit act as if it is reading the first symbol of an input sequence, allowing it to forget the previously computed state.

First, the one-hot vectors of title words  $T = [t_1 \dots t_n]$  are embedded individually to word level feature vectors  $T^w = [t_1^w \dots t_n^w]$ . To compute the n-gram level features, as in [37], we applied 1D convolution on the word embedding vectors. For the  $n$ th word, the convolution output with window size  $s$  is given by

$$\hat{t}_{s,n}^p = \tanh(W_c^s t_{n:n+s-1}^w), s \in 1, 2 \quad (10)$$

where  $W_c^s$  is the weight parameter. Max pooling was applied to obtain the final phrase level features. Then, the final phrase level features were encoded by Bi-GRU to obtain the title level feature representation  $T^t = [\hat{t}_1^p \dots \hat{t}_n^p]$ .

### 3.3. Emotion Category Feature Representation

When the data was collected, the final class was determined using the percentage of items that were predominantly labeled with a given emotion. The list of emotion categories is shown in Figure 2. As the data was provided with the percentage of each 20-emotion category, we considered them as an input in the training process. We used a three layer feed forward neural network to learn the feature vector from the emotion category  $C$ .

Polarity	Emotion Category	Abbreviation
<i>Positive</i>	<b>gratitude</b> , thankfulness, or indebtedness	grat
	<b>happiness</b> , calmness, pleasure, or ecstasy	happ
	<b>humility</b> , modesty, unpretentiousness, or simplicity	humi
	<b>love</b> or affection	love
	<b>optimism</b> , hopefulness, or confidence	opti
	<b>trust</b> , admiration, respect, dignity, or honor	trus
<i>Negative</i>	<b>anger</b> , annoyance, or rage	ange
	<b>arrogance</b> , vanity, hubris, or conceit	arro
	<b>disgust</b> , dislike, indifference, or hate	disg
	<b>fear</b> , anxiety, vulnerability, or terror	fear
	<b>pessimism</b> , cynicism, or lack of confidence	pessi
	<b>regret</b> , guilt, or remorse	regr
	<b>sadness</b> , pensiveness, loneliness, or grief	sadn
	<b>shame</b> , humiliation, or disgrace	sham
<i>Other or Mixed</i>	<b>agreeableness</b> , acceptance, submission, or compliance	agre
	<b>anticipation</b> , interest, curiosity, suspicion, or vigilance	anti
	<b>disagreeableness</b> , defiance, conflict, or strife	disa
	<b>surprise</b> , surrealism, amazement, or confusion	surp
	<b>shyness</b> , self-consciousness, reserve, or reticence	shyn
	<b>neutral</b>	neut

**Figure 2.** The list of emotions provided to annotators to label the title and art [1].

### 3.4. Co-Attention Layer

In the co-attention layer, attention mechanism we sequentially alternate between the generation of image, title, and category attentions consisting, briefly, of five steps. Starting from the encoded title/image/emotion category features, the proposed co-attention approach sequentially generates attention weights for each feature type, using the other two modalities as guides.

Specifically, we define an attention operation [36,37]  $\tilde{x} = A(X; g_1; g_2)$  that takes the image or title or category feature  $X$  and attention guidance  $g_1$  and  $g_2$  derived from title and image; title and category; or category and title as inputs and outputs the attended image, title or category vector. The operation can be expressed in the following steps:

$$\begin{aligned} H_i &= \tanh(W_x x_i + W_{g1} g_1 + W_{g2} g_2) \\ a_i &= \text{softmax}(w^T H_i), i = 1 \dots N \\ \tilde{x} &= \sum_{i=n}^N a_i x_i \end{aligned} \quad (11)$$

where  $X = [x_1; \dots; x_N] \in R^{d \times N}$  is the input sequence, and the fixed-length vectors  $g_1, g_2 \in R^d$  are attention guidance.  $W_x, W_{g1}, W_{g2} \in R^{h \times d}$  and  $w \in R^h$  are the embedding parameters to be learned.  $a$  is the attention weights of the input feature  $X$  and the weighted sum  $\tilde{x}$  is the weighted feature [36].

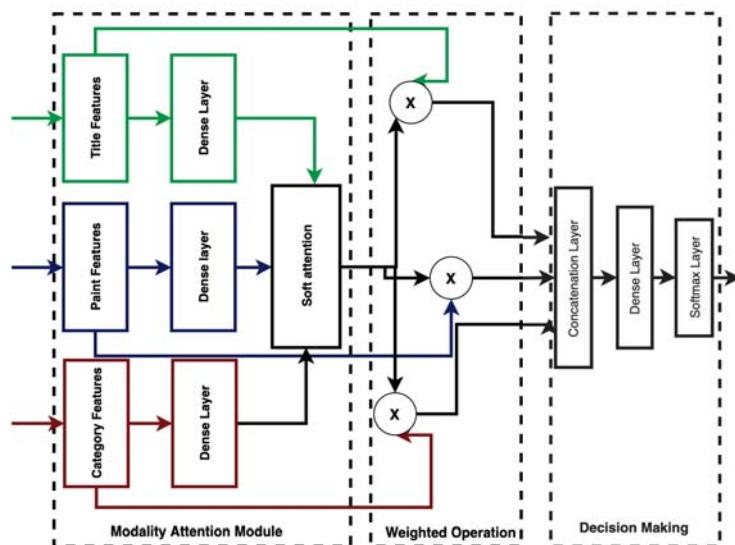
In the proposed sequential co-attention approach, the encoded title/category/image features are sequentially fed as input sequences to the attention module and the weighted features from the previous two steps are used as guidance [34,36,37]. First, the title features are summarized without guidance ( $\tilde{f}_0 = \text{Atten}(T; 0; 0)$ ) and secondly, the category features are weighted based on the summarized title features ( $\tilde{c}_0 = \text{Atten}(C; \tilde{f}_0; 0)$ ).

After that, the weighted image features will be computed using the weighted emotion category features ( $\tilde{c}_0$ ) and the title features  $\tilde{f}_0$  as guidance ( $\tilde{p} = \text{Atten}(P; \tilde{f}_0; \tilde{c}_0)$ ). In step 4 ( $\tilde{f} = \text{Atten}(T; \tilde{p}; \tilde{c}_0)$ ) and step 5 ( $\tilde{c} = \text{Atten}(C; \tilde{p}; \tilde{f})$ ), the title and category features will

also be re-weighted based on the results of the previous steps [36]. Finally, the weighted title/category/image features ( $\tilde{t}, \tilde{c}, \tilde{p}$ ) are further used for emotion prediction.

### 3.5. Weighted Modality Fusion

Decision-level fusion is a commonly used strategy for fusing heterogeneous inputs by combining the independent modality outputs using several specific rules [34]. However, the lack of mutual association learning across modalities is a major limitation in the application of decision-level fusion [40]. We used modality attention fusion, which enables feature-level system fusion and applies weighted modality scores across the extracted features to indicate the importance of different modalities. This preserves the advantages of both feature-level fusion and decision-level fusion [40]. The feature vector for each modality is first transformed into a fixed-length form. A three-layer feed-forward neural network (FFNN) was used to compute the attention weights for each modality, which were then used in the weighted average of the transformed feature vectors, as shown in Figure 3. The result is a single vector of fixed length.



**Figure 3.** Weighted modality fusion.

First, we implemented a three-layer feed-forward neural network to fuse the modality-specific features, and then we used softmax to generate the weighted score ( $s$ ) for the given modality as follows:

$$\begin{aligned} f &= \tanh(W_f[V_t, V_p, V_c] + b_f) \\ s &= \text{softmax}(f) \end{aligned} \quad (12)$$

where  $W_f$  and  $b + f$  are the trainable fusion parameters,  $s$  is an  $n$ -dimensional vector, and  $n = 3$  in this experiment (representing the modalities title, paint and category, respectively). We computed soft attention over the original modality features and concatenated them as in [34,40]. A dense layer was used to learn the associations over weighted modality-specific features by:

$$r = \tanh(W_r[s_t \tilde{t}, s_p \tilde{p}, s_c \tilde{c}]) \quad (13)$$

where  $r$  is the final fused representation, and  $W_r$  and  $b_r$  are the additional parameters for the final dense layer. We made the final decision by a softmax classifier using  $r$  as input.

### 3.6. Classification Layer

A three-layer fully connected neural network is used as the classification layer. The activation function of the hidden layer and the output layer are Relu and softmax functions, respectively. The loss function used is the categorical cross-entropy.

## 4. Experiment and Results

### 4.1. Dataset

Mohammad and Kiritchenko [1] created the WikiArt Emotions Dataset which includes emotion annotations for more than 4000 pieces of art from four Western styles (modern art, post-Renaissance art, Renaissance Art and Contemporary Art) and 22 style categories. The art is annotated via crowd sourcing for one or more of the twenty emotion categories. The final result of closely related emotion sets were arranged in three sets, such that “positive”, “negative” and “mixed or other”, as shown in Table 1.

**Table 1.** Main characteristics of the dataset used in the experiment.

Polarity	Emotion Category	Instances
Positive	gratitude, happiness, humility, love, optimism, trust	2578
Negative	anger, arrogance, disgust, fear, pessimism, regret, sadness, shame	838
Other or Mixed	agreeableness, anticipation, disagreeableness, surprise, shyness, neutral	689

### 4.2. Training Details

We implemented our proposed approach in Keras using the Tensorflow backend. The pre-trained ResNet model available in Keras is used for images, and the Glove word embedding program [41] for text was used to extract row feature vectors. The parameters of the pre-trained ResNet model and the parameters of the word embeddings were set during training. The Adam optimizer was used to optimize the loss function. The best hyper-parameters are listed in Table 2. In total, 70% of the data were used as the training set, 10% as the validation set and 20% as the test set.

**Table 2.** The best performing hyper-parameters used for the neural networks were determined by using a grid search [3].

Hyper-Parameters	Values
ResNet FC size	512
Batch size	32
Number of GRU hidden units	128
Dropout rate for GRU	0.4
Number of epochs	40
Learning rate	0.001
Word embedding dimensions	100

### 4.3. Baselines

- Bi-LSTM (Text Only): Bi-LSTM is one of the most popular methods for addressing many text classification problems. It leverages a bidirectional LSTM network for learning text representations and then uses a classification layer to make a prediction.
- CNN (Image Only): CNN with six hidden layers was implemented. The first two convolutional layers contain 32 kernels of size  $3 \times 3$  and the second two convolutional layers have 64 kernels of size  $3 \times 3$ . The second and fourth convolutional layers are

interleaved with max-pooling layers of dimension  $2 \times 2$  with a dropout of 0.3. Then, a fully connected layer with 256 neurons and a dropout of 0.4 is followed.

- Multimodal approaches (text and image): two multimodal approaches, namely Resnet\_GRU without attention and Resnet\_GRU attention from the previous work [3], in the same task were also implemented.

#### 4.4. Results and Discussion

The proposed approach was compared with the three unimodal baseline approaches and three multimodal approaches. As shown in Table 3, the proposed model improves the unimodal-based methods, which use only a single feature type, and the multimodal models, which use only information from the image and title modalities.

The proposed approach gained 8.4%, 9% and 11.5% in terms of accuracy when compared to the unimodal text-based, emotion category and image-based networks, respectively. These significant improvements confirm the importance of extracting and using information from different modalities in human emotion recognition and analysis.

**Table 3.** Performance on test set in terms of the accuracy on the three polarities.

Model	Channel	Accuracy	Loss
CNN	Image	0.683	0.663
Bi-LSTM	Title	0.658	0.810
FFNN	Category	0.689	0.441
ResNet_GRU without attention	Paint, title	0.713	0.710
ResNet_GRU with attention	Paint, title	0.741	0.130
Our new model with concatenation	Paint, title and category	0.724	0.684
Our new model	Paint, title and category	0.773	0.143

Furthermore, we compared the proposed approach that uses information from the three modalities with two multimodal approaches that use information from image and title modalities, namely Resnet\_GRU without attention and Resnet\_GRU with attention. Our proposed approach has outperformed Resnet\_GRU without attention by 6% and Resnet\_GRU with attention by 3.2%.

Our proposed approach uses information from the three modalities which are image, title, and emotion category, but emotion category values are used during the training phase only. The main reason for using the emotion category during the training as one of the modality inputs is to help the model learn from expert knowledge and to see the impact of expert knowledge on model training. The experimental results have shown that using expert knowledge helped the model learn better, as shown in Table 3.

To show the advantage of weighted modality fusion over other fusion methods, we compared the weighted modality fusion model with other fusion methods. The experimental results showed that the proposed sequential-based co-attention feature learning and weighted modality fusion approaches can learn better for different categories, which implies that using pre-trained models with sequential attention and weighted modality fusion is a reasonable choice for emotion recognition from art. Figure 4 shows how our model learns from the training dataset and the generalizability of our model on the validation set, which confirms that the chosen model perfectly fits to address the emotion recognition tasks.

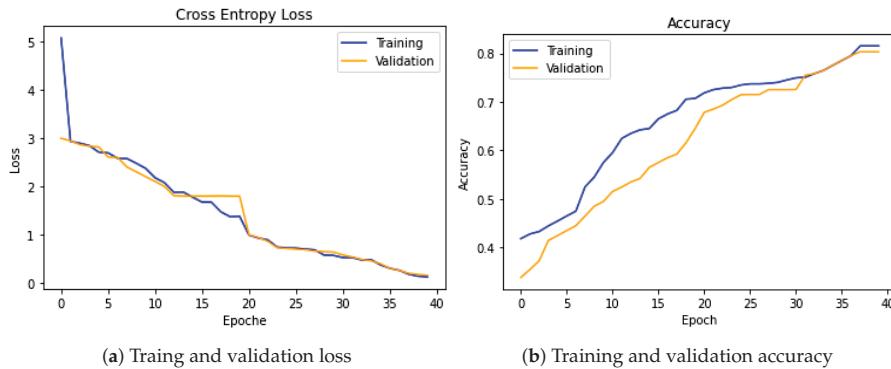


Figure 4. Cross-entropy loss and accuracy during the training and validation steps are shown in (a,b), respectively.

## 5. Conclusions

In this work, we proposed sequential-based attention to extract features from three modalities (title, art, and emotion category) and a weighted fusion approach to fuse the three modalities in the decision process. Our system used feature attention (sequential co-attention) and modality attention (weighted fusion) to select the representative information at both feature and modality levels. The experimental results on the WikiArt dataset demonstrated the effectiveness of the proposed model and the usefulness of the three modalities. Although our model was evaluated for emotion recognition in art, it can potentially be applied to other similar tasks involving different modalities.

**Author Contributions:** Conceptualization, T.M.T. and T.H.; methodology, T.M.T.; software T.M.T. and S.H.; validation, T.M.T. and S.H.; formal analysis; data curation T.M.T.; writing—original draft preparation, T.M.T.; writing—review and editing T.M.T., S.H. and T.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are openly available at WikiArt Emotions Project webpage: <http://saifmohammad.com/WebPages/wikiartemotions.html> (accessed on 16 August 2021).

**Acknowledgments:** This research is supported by the ÚNKP-20-4 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund. Supported by Telekom Innovation Laboratories (T-Labs), the Research and Development unit of Deutsche Telekom. Project no. ED\_18-1-2019-0030 (Application domain specific highly reliable IT solutions subprogramme) has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the Thematic Excellence Programme funding scheme.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Mohammad, S.; Kiritchenko, S. WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 7–12 May 2018.
- Tripathi, S.; Beigi, H.S.M. Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning. *arXiv* **2018**, arXiv:1804.05788.

3. Tashu, T.M.; Horváth, T. Attention-Based Multi-modal Emotion Recognition from Art. Pattern Recognition. In *Proceedings of the ICPR International Workshops and Challenges, Virtual Event, 10–15 January 2021; Part III*; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; pp. 604–612.
4. Sreeshakthy, M.; Preethi, J. Classification of Human Emotion from Deep EEG Signal Using Hybrid Improved Neural Networks with Cuckoo Search. *BRAIN Broad Res. Artif. Intell. Neurosci.* **2016**, *6*, 60–73.
5. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [\[CrossRef\]](#)
6. Clavel, C.; Vasilescu, I.; Devillers, L.; Richard, G.; Ehrette, T. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Commun.* **2008**, *50*, 487–503. [\[CrossRef\]](#)
7. Khalfallah, J.; Slama, J.B.H. Facial Expression Recognition for Intelligent Tutoring Systems in Remote Laboratories Platform. *Procedia Comput. Sci.* **2015**, *73*, 274–281. [\[CrossRef\]](#)
8. Knapp, R.B.; Kim, J.; André, E. Physiological Signals and Their Use in Augmenting Emotion Recognition for Human–Machine Interaction. In *Emotion-Oriented Systems: The Humaine Handbook*; Cowie, R., Pelachaud, C., Petta, P., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 133–159. [\[CrossRef\]](#)
9. Shenoy, A.; Sardana, A. Multilogue-Net: A Context-Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*; Association for Computational Linguistics: Seattle, WA, USA, 2020; pp. 19–28. [\[CrossRef\]](#)
10. Yoon, S.; Dey, S.; Lee, H.; Jung, K. Attentive Modality Hopping Mechanism for Speech Emotion Recognition. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3362–3366. [\[CrossRef\]](#)
11. Liu, G.; Yan, Y.; Ricci, E.; Yang, Y.; Han, Y.; Winkler, S.; Sebe, N. *Inferring Painting Style with Multi-Task Dictionary Learning*; AAAI Press: Cambridge, MA, USA, 2015; pp. 2162–2168.
12. Wang, Y.; Takatsuka, M. SOM based artistic styles visualization. In Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME), San Jose, CA, USA, 15–19 July 2013; pp. 1–6.
13. Szegedy, C.; Wei, L.; Yangqing, J.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
14. Sartori, A.; Culibrk, D.; Yan, Y.; Sebe, N. *Who's Afraid of Itten: Using the Art Theory of Color Combination to Analyze Emotions in Abstract Paintings (MM '15)*; Association for Computing Machinery: New York, NY, USA, 2015; pp. 311–320. [\[CrossRef\]](#)
15. Zhao, S.; Gao, Y.; Jiang, X.; Yao, H.; Chua, T.S.; Sun, X. *Exploring Principles-of-Art Features For Image Emotion Recognition*; Association for Computing Machinery: New York, NY, USA, 2014; pp. 47–56. [\[CrossRef\]](#)
16. Yanulevskaya, V.; van Gemert, J.C.; Roth, K.; Herbold, A.K.; Sebe, N.; Geusebroek, J.M. Emotional valence categorization using holistic image features. In Proceedings of the 2008 15th IEEE International Conference on Image Processing, San Diego, CA, USA, 12–15 October 2008; pp. 101–104.
17. Scherer, K.; Johnstone, T.; Klasmeyer, G. *Handbook of Affective Sciences—Vocal Expression of Emotion*; Oxford University: Oxford, UK, 2003; pp. 433–456.
18. Navarretta, C. *Individuality in Communicative Bodily Behaviours*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 417–423. [\[CrossRef\]](#)
19. Seyeditabar, A.; Tabari, N.; Gholizadeh, S.; Zadrozny, W. Emotion Detection in Text: Focusing on Latent Representation. *arXiv* **2019**, arXiv:abs/1907.09369.
20. Yeh, S.L.; Lin, Y.S.; Lee, C.C. An Interaction-aware Attention Network for Speech Emotion Recognition in Spoken Dialogs. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6685–6689. [\[CrossRef\]](#)
21. Castellano, G.; Kessous, L.; Caridakis, G. Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech. In *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*; Peter, C., Beale, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 92–103.
22. Sikka, K.; Dykstra, K.; Sathyanarayana, S.; Littlewort, G.; Bartlett, M. *Multiple Kernel Learning for Emotion Recognition in the Wild*; Association for Computing Machinery: New York, NY, USA, 2013; pp. 517–524. [\[CrossRef\]](#)
23. Kim, Y.; Lee, H.; Provost, E.M. Deep learning for robust feature generation in audiovisual emotion recognition. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 3687–3691.
24. Majumder, N.; Hazarika, D.; Gelbukh, A.; Cambria, E.; Poria, S. Multimodal Sentiment Analysis Using Hierarchical fusion with context modeling. *Knowl. Based Syst.* **2018**, *161*, 124–133. [\[CrossRef\]](#)
25. Ren, M.; Nie, W.; Liu, A.; Su, Y. Multi-modal Correlated Network for emotion recognition in speech. *Vis. Inform.* **2019**, *3*, 150–155. [\[CrossRef\]](#)
26. Yoon, S.; Byun, S.; Dey, S.; Jung, K. Speech Emotion Recognition Using Multi-hop Attention Mechanism. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2822–2826.

27. Lian, Z.; Li, Y.; Tao, J.; Huang, J. Investigation of Multimodal Features, Classifiers and Fusion Methods for Emotion Recognition. *arXiv* **2018**, arXiv:1809.06225.
28. Pan, Z.; Luo, Z.; Yang, J.; Li, H. Multi-Modal Attention for Speech Emotion Recognition, 2020. Available online: <http://xxx.lanl.gov/abs/2009.04107> (accessed on 16 August 2021).
29. Siriwardhana, S.; Reis, A.; Weerasekera, R.; Nanayakkara, S. Jointly Fine-Tuning “BERT-like” Self Supervised Models to Improve Multimodal Speech Emotion Recognition. *arXiv* **2020**, arXiv:2008.06682.
30. Liu, G.; Tan, Z. Research on Multi-modal Music Emotion Classification Based on Audio and Lyric. In Proceedings of the 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 12–14 June 2020; Volume 1, pp. 2331–2335. [CrossRef]
31. Machajdik, J.; Hanbury, A. *Affective Image Classification Using Features Inspired by Psychology and Art Theory*; Association for Computing Machinery: New York, NY, USA, 2010; pp. 83–92. [CrossRef]
32. Yanulevskaya, V.; Uijlings, J.; Bruni, E.; Sartori, A.; Zamboni, E.; Bacci, F.; Melcher, D.; Sebe, N. *In the Eye of the Beholder: Employing Statistical Analysis and Eye Tracking for Analyzing Abstract Paintings*; Association for Computing Machinery: New York, NY, USA, 2012; pp. 349–358. [CrossRef]
33. Sartori, A.; Yan, Y.; Özbal, G.; Almila, A.; Salah, A.; Salah, A.A.; Sebe, N. *Looking at Mondrian’s Victory Boogie-Woogie: What Do I Feel*; AAAI Press: Cambridge, MA, USA, 2015; pp. 2503–2509.
34. Cai, Y.; Cai, H.; Wan, X. *Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 2506–2515. [CrossRef]
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
36. Wang, P.; Wu, Q.; Shen, C.; van den Hengel, A. The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Questions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3909–3918.
37. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS’16)*; Curran Associates Inc.: Red Hook, NY, USA, 2016; pp. 289–297.
38. Chung, J.; Gülcabay, Ç.; Cho, K.; Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* **2014**, arXiv:1412.3555.
39. Tashu, T.M. Off-Topic Essay Detection Using C-BGRU Siamese. In Proceedings of the 2020 IEEE 14th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 3–5 February 2020; pp. 221–225. [CrossRef]
40. Gu, Y.; Yang, K.; Fu, S.; Chen, S.; Li, X.; Marsic, I. Hybrid Attention based Multimodal Network for Spoken Language Classification. In *Proceedings of the 27th International Conference on Computational Linguistics*; Association for Computational Linguistics: Santa Fe, NM, USA, 2018; pp. 2379–2390.
41. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1532–1543. [CrossRef]



Article

# Towards Generating and Evaluating Iconographic Image Captions of Artworks

Eva Cetinic <sup>1,2</sup>

<sup>1</sup> Rudjer Boskovic Institute, Bijenicka Cesta 54, 10000 Zagreb, Croatia; [ecetinic@irb.hr](mailto:ecetinic@irb.hr)

<sup>2</sup> Department of Computer Science, Durham University, Durham DH1 3LE, UK

**Abstract:** To automatically generate accurate and meaningful textual descriptions of images is an ongoing research challenge. Recently, a lot of progress has been made by adopting multimodal deep learning approaches for integrating vision and language. However, the task of developing image captioning models is most commonly addressed using datasets of natural images, while not many contributions have been made in the domain of artwork images. One of the main reasons for that is the lack of large-scale art datasets of adequate image-text pairs. Another reason is the fact that generating accurate descriptions of artwork images is particularly challenging because descriptions of artworks are more complex and can include multiple levels of interpretation. It is therefore also especially difficult to effectively evaluate generated captions of artwork images. The aim of this work is to address some of those challenges by utilizing a large-scale dataset of artwork images annotated with concepts from the Iconclass classification system. Using this dataset, a captioning model is developed by fine-tuning a transformer-based vision-language pretrained model. Due to the complex relations between image and text pairs in the domain of artwork images, the generated captions are evaluated using several quantitative and qualitative approaches. The performance is assessed using standard image captioning metrics and a recently introduced reference-free metric. The quality of the generated captions and the model's capacity to generalize to new data is explored by employing the model to another art dataset to compare the relation between commonly generated captions and the genre of artworks. The overall results suggest that the model can generate meaningful captions that indicate a stronger relevance to the art historical context, particularly in comparison to captions obtained from models trained only on natural image datasets.



**Citation:** Cetinic, E. Towards Generating and Evaluating Iconographic Image Captions of Artworks. *J. Imaging* **2021**, *7*, 123. <https://doi.org/10.3390/jimaging7080123>

Academic Editors: Giovanna Castellano, Gennaro Vessio and Fabio Bellavia

Received: 7 June 2021

Accepted: 19 July 2021

Published: 23 July 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Image captioning refers to the task of generating a short text that describes the content of an image based only on the image input. This usually implies recognizing objects and their relationships in an image. Those descriptions should be meaningful and accurate in relation to the image content. In resolving this task, significant progress has recently been made using multimodal deep learning models. However, most of the research in this field is performed on datasets of natural images, while the specific aspects of generating captions for artwork images have not yet been systematically explored.

A common prerequisite for training deep neural captioning models are large datasets of semantically related image and sentence pairs. In the domain of natural images, several well-known large-scale datasets are commonly used for this task, such as the MS COCO [1], Flickr30 [2] and Visual Genome [3] dataset. The availability of such large datasets enabled the development of image captioning models that achieve impressive results in generating high quality captions for photographs of various objects and scenes. However, the task of generating image captions still remains difficult for domain-specific image collections. In particular, in the context of visual art and cultural heritage, generating image captions is an open problem with various challenges. The lack of a truly large-scale dataset of artwork

images paired with adequate descriptions represents one of the major difficulties. Furthermore, it is important to address what kind of description would be regarded as “adequate” in the context of art historical data collections. Taking into account Erwin Panofsky’s three levels of analysis [4], we can distinguish the “pre-iconographic” description, “iconographic” description and the “iconologic” interpretation as possibilities of aligning meaningful, yet very different textual descriptions with the same image. Image captioning in the context of natural images is usually performed at the level of “pre-iconographic” descriptions, which implies simply describing the content and listing the objects that are depicted in an image. For artwork images this type of description represents only the most basic level of visual understanding and is not considered to be particularly useful for performing multimodal analysis and retrieval within art collections.

A more interesting, as well as more challenging, task would be to generate “iconographic” captions that describe the contextual aspect of the subject matter. Creating a dataset for such a complex task is difficult because it requires expert knowledge in the process of collecting sentence-based descriptions of images. Several such art datasets of image-text pairs exist, but those mostly consist of only a few thousand examples and are therefore not suitable for training deep neural network models in the current state-of-the-art setting for image captioning. However, there are several existing large-scale artwork collections that associate images with textual descriptions in the form of keywords and specific concepts. In particular, a large-scale artwork dataset, published under the name “Iconclass AI Test Set” [5], represents a collection of various artwork images assigned with alphanumeric classification codes that correspond to notations from the Iconclass system [6]. Iconclass is a classification system designed for art and iconography and is widely accepted by museums and art institutions as a tool for the description and retrieval of subjects represented in images. The idea of this work is to use a concatenation of the various code descriptions associated with an image as textual inputs for training an image captioning model. Although the “Iconclass AI Test Set” is not structured primarily as an image captioning dataset, each code is paired with its “textual correlate”—a description of the iconographic subject of the particular Iconclass notation. The first methodological step of the approach presented in this work includes extracting and preprocessing the given annotations into clean textual description and creating the “Iconclass Caption” dataset. This dataset is then used to fine-tune a pretrained unified vision-language model on the down-stream task of image captioning [7]. Transformer-based vision-language pretrained models currently represent the leading approach in solving a variety of tasks in the intersection of computer vision and natural language processing.

The work presented in this paper is an extension of a previous work that represents one of the first attempts in generating captions for artworks [8]. The methodological approach is similar and the additional contribution of this paper is primarily focused on the problem of evaluating the generated image captions. The previous work showed that standard reference-based image metrics are not very suitable for assessing the quality of image captions because they take into account only the relation between the generated and ground-truth caption, and not the relation between the caption and the image itself, which is particularly important in the context of artworks. Recently, significant advances have been achieved in transforming image and text embeddings into a joint feature space. Based on those findings, this work additionally explores how CLIP (Contrastive Language-Image Pre-training), a newly introduced cross-modal model pretrained on very large dataset of 400 M image+text pairs extracted from the web [9], and reference-free captioning metrics defined based on CLIP features [10], can be used to evaluate the generated iconographic captions.

## 2. Related Work

The availability of large collections of digitized artwork images fostered research initiatives in the intersection of artificial intelligence and art history. Most commonly, research in this area focuses on addressing problems related to computer vision in the context of art data, such as image classification [11–13], visual link retrieval [14–16], object

and face detection [17,18], pose and character matching [19,20], analysis of visual patterns and conceptual features [21–24], and computational aesthetics [25–27]. A comprehensive overview of research activities in this area can be found in several survey papers [28–30].

Recently, there has been a surge of interest in topics related to jointly exploring both visual and textual modalities of artwork collections. Pioneering works in this research area addressed the task of multimodal retrieval. In particular, Ref. [31] introduced the SemArt dataset, a collection of fine-art images associated with textual comments, with the aim to map the images and their descriptions in a joint semantic space. They compare different combinations of visual and textual encodings, as well as different methods of multimodal transformation. In projecting the visual and textual encodings in a common multimodal space, they achieve the best results by applying a neural network trained with cosine margin loss on ResNet50 features as visual encodings and bag of word as textual encodings. The task of creating a shared embedding space was also addressed in [32], where the authors introduce a new visual semantic dataset named BibleVSA, a collection of miniature illustrations and commentary text pairs, and explore supervised and semi-supervised approaches to learning cross-references between textual and visual information in documents. In [33], the authors present the Artpedia dataset, consisting of 2930 images annotated with visual and contextual sentences. They introduce a cross-modal retrieval model that projects images and sentences in a common embedding space and discriminates between contextual and visual sentences of the same image. A similar extension of this approach to other artistic datasets was presented in [34]. Recently, Banar et al. introduced a study that explores how Iconclass codes can be automatically assigned to visual artworks using a cross-modal retrieval set-up [35].

Apart from multimodal retrieval, another recently emerging topic of interest is visual question answering (VAQ). In [36], the authors annotated a subset of the ArtPedia dataset with visual and contextual question-answer pairs and introduced a question classifier that discriminates between visual and contextual questions and a model that is able to answer both types of questions. In [37], the authors introduce a novel dataset AQUA (Art QUEStion Answering), which consists of automatically generated visual and knowledge-based question-answer pairs, and also present a two-branch model where the visual and knowledge questions are handled independently.

The task of image captioning has not been significantly studied in the context of art images. A limited number of studies contributed to the task of generating descriptions of artwork images using deep neural networks. For example, Ref. [38] proposes an encoder-decoder framework for generating captions of artwork images where the encoder (ResNet18 model) extracts the input image feature representation and the artwork type representation, while the decoder is a long short-term memory (LSTM) network. They introduce two image captioning datasets referring to ancient Egyptian art and ancient Chinese art, which contain 17,940 and 7607 images, respectively. Another work [39] presented a novel captioning dataset for art historical images consisting of 4000 images across nine iconographies, along with a description for each image consisting of one or more paragraphs. They used this dataset to fine-tune different variations of image captioning models based on the well-known encoder-decoder approach introduced in [40]. As already mentioned, this paper represent an extension of the image captioning approach presented in [8].

Motivated by the success of utilizing large-scale pretrained language models such as the BERT (Bidirectional Encoder Representations from Transformers) model [41] for different tasks related to natural language processing, recently significant research progress has been made by adopting transformer-based models for a variety of multimodal tasks. Transformer-based vision-language models are designed to learn joint representations that combine and align information from both modalities. It has been shown that models pretrained on intermediate tasks with unsupervised learning objectives using large datasets of image-text pairs achieve remarkable results when applied to different down-stream tasks such as image captioning, cross-modal retrieval or visual question answering [7,42–44]. Furthermore, recently an efficient method of learning from natural language supervision

was introduced as the CLIP (Contrastive Language-Image Pre-training) model [9]. The model is a result of training an image and text encoder to predict the correct pairs of image-text training examples using large amounts of publicly available internet data. The CLIP model showed very promising results on a variety of image-text similarity estimation tasks and was recently introduced as a novel way of establishing a reference-free image captioning metric [10]. This paper explores how those newly introduced image captioning metrics, as well as CLIP image and text representations, can be used to evaluate captions in the context of artworks.

### 3. Methodology

#### 3.1. Datasets

##### 3.1.1. Iconclass Caption Dataset for Training and Evaluation

The main dataset used in this work is the “Iconclass AI Test Set” [5] dataset. The dataset contains, in total, 87,749 images, and in this work 86,530 valid image-text pairs are used for training and evaluating the image captioning model (1219 images do not have valid codes/textual notations assigned to them). The dataset includes a very diverse collection of images sampled from the Arkyves database [www.arkyves.org](http://www.arkyves.org) (accessed on 21 June 2021). It includes images of various types of artworks such as paintings, posters, drawings, prints, manuscripts pages, etc. Each image is associated with one or more codes linked to labels from the Iconclass classification system. The authors of the “Iconclass AI Test Set” provide a json file with the list of images and corresponding codes, as well as an Iconclass Python package to perform analysis and extract information from the assigned classification codes. To extract textual descriptions of images for the purpose of this work, the English textual descriptions of each code associated with an image are concatenated. Further preprocessing of the descriptions includes removing text in brackets and some recurrent uppercased dataset-specific codes. In this dataset, the text in brackets most commonly includes very specific named entities, which are considered a noisy input in the image captioning task. Therefore, when preprocessing the textual items, all the text in brackets is removed, even at the cost of sometimes removing useful information.

Figure 1 shows several example images from the Iconclass Caption dataset and their corresponding descriptions before and after preprocessing. Depending on the number of codes associated with each image, the final textual descriptions can significantly vary in length. Additionally, due to the specific properties of this dataset, the image descriptions are not structured as sentences but as a list of comma-separated words and phrases.

The textual descriptions are represented as a concatenation of text phrases related to the Iconclass codes. One image in the dataset can be associated with one or more textual phrases. To better understand the configuration of the dataset, Figure 2 shows a distribution of the most commonly included textual phrases (Iconclass codes).

Due to this type of structure and having only one reference caption for each image, the Iconclass Caption dataset is not a standard image captioning dataset. However, having in mind the difficulties of obtaining adequate textual descriptions for images of artworks, this dataset can be considered as a valuable source of image-text pairs in the current context, particularly due to the large number of annotated images that enables training deep neural models. In the experimental setting, a subset of approximately 76,000 items is used for training the model, while around 5000 items are used for validation and 5000 for testing.



**Original description:** head turned to the right, wig, bookshelves, neck-gear: jabot, clothing for the upper part of the body (VEST) , party clothes, festive attire (+ men's clothes), quill, book, historical persons (portraits and scenes from the life) (+ (full) bust portrait),

**Clean description:** head turned to the right, wig, bookshelves, neck-gear: jabot, clothing for the upper part of the body , party clothes, festive attire , quill, book, historical persons .



**Original description:** (human) skull, bones in general (human body), death's head, skull (symbol of Death)

**Clean description:** skull, bones in general , death's head, skull.



**Original description:** plants and herbs (ARMORACIA RUSTICANA), plants and herbs (HORSERADISH), proverbs, sayings, etc. (IM GAUMEN), proverbs, sayings, etc. (DER BEISSENDE),

**Clean description:** plants and herbs , proverbs, sayings.



**Original description:** Mary standing (or half-length), the Christ-child sitting on her arm (Christ-child to Mary's left),

**Clean description:** Mary standing , the Christ-child sitting on her arm.



**Original description:** adult woman, manuscript of musical score, writer, poet, author (+ portrait, self-portrait of artist), pen, ink-well, paper (writing material), codex, inscription, historical events and situations (1567), historical person (MONTENAY, Georgette de) - BB - woman - historical person (MONTENAY, Georgette de) portrayed alone, proverbs, sayings, etc. (O PLUME EN LA MAIN NON Vaine)

**Clean description:** adult woman, manuscript of musical score, writer, poet, author , pen, ink-well, paper , codex, inscription, historical events and situations , historical person, woman - historical person portrayed alone, proverbs, sayings.

Figure 1. Example images from the Iconclass Caption dataset and their corresponding descriptions before and after preprocessing.

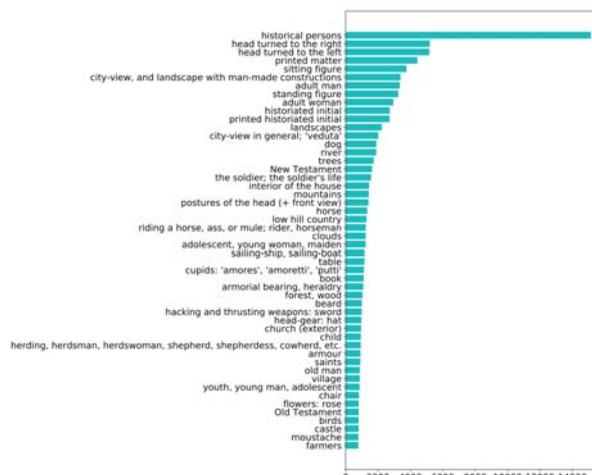


Figure 2. Distribution of textual descriptions in the Iconclass Caption dataset showing the 50 most commonly occurring words/phrases (Iconclass codes) in the whole dataset.

### 3.1.2. Wikiart Dataset for Evaluation

In order to explore how the proposed approach works on another artwork dataset, a subset of 52,562 images of paintings from the WikiArt, [www.wikiart.org](http://www.wikiart.org) (accessed on 1 February 2020), collection was used. Images in the WikiArt dataset are annotated with a broad set of labels (e.g., style, genre, artist, technique, date of creation, etc.); therefore, one aspect of the evaluation process includes analysing how the generated captions relate to genre labels because genre labels indicate the category of the subject matter that is depicted (e.g., portrait, landscape, religious paintings, etc.). Furthermore, this dataset is used to explore the difference between captions generated using a model trained on artwork images and models trained on natural image datasets.

### 3.2. Image Captioning Model

For the purpose of training an image captioning model, in this work the unified vision-language pretraining model (VLP) introduced in [7] was employed. This model is denoted as “unified” because the same pretrained model can be fine-tuned for different types of tasks. These tasks include both vision-language generation (e.g., image captioning) and vision-language understanding (e.g., visual question answering). The model is based on an encoder-decoder architecture comprised of 12 transformer blocks. The model input consist of image embedding, text embedding and three special tokens that indicate the start of the image input, the boundary between the visual and textual input and the end of the textual input. The image input consists of 100 object classification aware region features extracted using the Faster R-CNN (region-based convolutional neural networks) model [45] pretrained on the Visual Genome dataset [3]. For a more detailed description of the overall VLP framework and pretraining objectives, the reader is referred to [7]. The experiments introduced in this work employ, as the base model, the VLP model pretrained on the Conceptual Captions dataset [46] using the sequence-to-sequence objective. This base model is fine-tuned on the Iconclass Caption dataset using recommended fine-tuning configurations, namely training with a constant learning rate of 3e-5 for 30 epochs. The weights of the Iconclass fine-tuned model, together with the data used for training the model (image IDs and descriptions), are available here: <https://github.com/EvaCet/Iconclass-image-captioning> (accessed on 22 July 2021).

### 3.3. Evaluation of the Generated Captions

The evaluation of the model’s performance includes both quantitative and qualitative analyses of the generated captions. To quantitatively evaluate the generated captions, standard language evaluation metrics for image captioning and novel reference-free image captioning methods are used. The standard metrics include the four BLEU metrics [47], METEOR [48] ROUGE [49] and CIDEr [50]. BLUE, ROUGE and METEOR are metrics that originate from machine translation tasks, while CIDEr was specifically developed for image caption evaluation. The BLUE metrics represent n-gram precision scores multiplied by a brevity penalty factor to assess the length correspondence of candidate and reference sentences. ROUGE is a metric that measures the recall of n-grams and therefore rewards long sentences. Specifically, ROUGE-L measures the longest matching sequence of words between a pair of sentences. METEOR represents the harmonic mean of precision and recall of unigram matches between sentences and additionally includes synonyms and paraphrase matching. CIDEr measures the cosine similarity between TF-IDF weighted n-grams of the candidate and the reference sentences. The TF-IDF weighting of n-grams reduces the score of frequent n-grams and appoints higher scores to distinctive words.

As the standard image captioning metrics measure the relation between generated and original captions, they do not address the relation between the image itself and the generated caption. Although translating images and text in a joint semantic space has been an ongoing research topic, the recently introduced CLIP model [9] achieves significant performance improvements in assessing the similarity between image and text. Based on the advanced performance of this model, Hassel et al. [10] introduce a novel reference-free

metric called CLIPScore, which, according to their study, achieves the highest correlation with human judgements and outperforms existing reference-based metrics. The CLIPscore represents a rescaled value (multiplied by factor of 2.5) of the cosine similarity between image and generated caption text embeddings obtained using the CLIP ViT-B/32 model for feature extraction. They also introduce a reference-augmented version of this score, the RefCLIPScore, which is computed as a harmonic mean of the CLIPScore and the maximal reference cosine similarity. Image captioning datasets usually include more than one reference sentence per image; however, the Iconclass Caption dataset includes only one reference description. Therefore, in this work, the RefCLIPScore is described as a harmonic mean between the rescaled cosine similarity between the CLIP embeddings of the image and generated caption (the CLIPScore) and the value of the cosine similarity between the CLIP embeddings of the reference caption and generated caption.

## 4. Results and Discussion

### 4.1. Quantitative Results

The relation between the generated captions and the reference captions on the Iconclass Caption test set was evaluated using standard image captioning metrics. To evaluate the relation between the generated caption and the input image, the new CLIPScore metric was used, both in its original and reference-augmented versions. The results on the Iconclass Caption test set are presented in Table 1. The Iconclass Caption test set contains 5192 images, but the reported CLIP-S and RefCLIP-S values are calculated only on a subset of 4928 images where the generated captions are shorter than 76 tokens, together with tokens that indicate the end and beginning of the text sequence. This was carried out because the CLIP model, which serves as a basis for the CLIPScore metric, was trained with the maximal textual sequence length set at 76 tokens. As the Iconclass Caption dataset contains descriptions of various lengths, including very long ones, some of the generated captions are also long. In order to test the model on all the examples in the Iconclass Caption test set, an alternative version of the whole dataset was created where all image descriptions have been shortened in order to fit into the range of the maximal sequence length. As most of the descriptions consist of comma-separated concatenations of words and phrases, the shortening has been performed to keep only so many concatenated phrases to meet the 76 tokens limit. However, this shortening of the descriptions led to an overall deterioration of the captioning results in comparison with the results on the original, non-shortened dataset presented in Table 1 (the values of the metric scores on the alternative version of the dataset are: Bleu 1: 0.11; Bleu 2: 0.10; Bleu 3: 0.092; Bleu 4: 0.08; METEOR: 0.115; ROUGE-L: 0.302; CIDEr: 1.57; CLIP-S: 0.596; RefCLIP: 0.677). It was therefore decided to present and use the model trained on the original version of the dataset for further analysis and to report the CLIP-S and RefCLIP-S scores on a slightly smaller subset of the test set.

**Table 1.** Values of the evaluation metrics used for assessing the performance of the iconographic image captioning model on the Iconclass Caption test set. \*CLIP-S and RefCLIP-S values are reported on a subset of the test set.

Evaluation Metric	Value ( $\times 100$ )
BLEU 1	14.8
BLEU 2	12.8
BLEU 3	11.3
BLEU 4	10.0
METEOR	11.7
ROUGE-L	31.9
CIDEr	172.1
CLIP-S *	59.67
RefCLIP-S *	68.35

The current results cannot be compared with any other work because the experiments were performed on a new and syntactically and semantically different dataset. However, the quantitative evaluation results are included to serve as a benchmark for future work. In comparison with current state-of-the-art caption evaluation results on natural image datasets (e.g., BLEU4  $\approx$  37 for MS COCO and  $\approx$ 30 for Flickr30 datasets) [7,51], the BLEU scores are lower for the Iconclass dataset. A similar behaviour was also reported in another study addressing iconographic image captioning [39]. On the other hand, the CIDEr score is quite high in comparison to the one reported for natural image datasets (e.g., CIDEr  $\approx$  116 for MS COCO and  $\approx$ 68 for Flickr30 dataset) [7,51]. To better understand how standard metrics relate to the novel metrics, Table 2 shows the Spearman’s rank correlation coefficient between the values of standard and novel captioning metrics on the Iconclass Caption test set.

**Table 2.** Spearman’s correlation coefficients between the values of standard and new metric scores on the Iconclass Caption test set ( $p$ -value  $< 0.001$ ).

Standard Metric	Correlation with CLIP-S	Correlation with Ref CLIP-S
BLEU-1	0.355	0.686
BLEU-2	0.314	0.647
BLEU-3	0.281	0.629
BLEU-4	0.236	0.602
METEOR	0.315	0.669
ROUGE-L	0.298	0.647
CIDEr	0.315	0.656

It is questionable how adequate standard reference-based metrics are in assessing the overall quality of the captions in this particular context because they mostly measure the word overlap between generated and reference captions. They are not designed to capture the semantic meaning of a sentence and therefore it is particularly difficult to evaluate iconographic descriptions. Furthermore, they are not appropriate for measuring very short descriptions which are quite common in the IconClass Caption dataset. Moreover, because they do not address the relation between the generated caption and the image content, the standard image captioning score could be low even if the generated caption is semantically aligned with the image content. In Figure 3, several such examples from the Iconclass Caption test set are presented, together with the values of the standard and new metrics.

In some examples within the Iconclass dataset, the generated caption is even more related to the image content than the ground-truth description (example image in row 3 in Figure 3) and interestingly the CLIP-Score is, in this case, higher than the usually higher RefCLIP-Score. Furthermore, those examples indicate that the standard evaluation metrics are not very suitable in assessing the relevance of generated captions for this particular dataset. Therefore, a qualitative analysis of the results is also required in order to better understand potential contributions and drawbacks of the proposed approach.



**GT:** ‘Oriente’ , wig, interior of the house, table, chair, table-cloth, pipe tobacco, head-gear: hat, head-gear , neck-gear: jabot, sewing, marriage, married couple, ‘matrimonium’ , pen, ink-well, book.

**Caption:** sitting figure , head turned to the left , head turned to the right, adult man , adult woman , historical persons .

BLEU 1: 0.03765	ROUGE: 0.0451
BLEU 2: $1.22 \times 10^{-9}$	CIDEr: 0.0041
BLEU 3: $3.99 \times 10^{-12}$	CLIP-S: 0.702
BLEU 4: $2.31 \times 10^{-13}$	RefCLIP-S: 0.689
METEOR: 0.035	

**Figure 3. Cont.**



**GT:** church , water course, city-view, and landscape with man-made constructions, street, clouds, rowing-boat, canoe, sailing-ship, sailing-boat, windmill, cow.

**Caption:** ships .

BLEU 1:  $1.12 \times 10^{-22}$  ROUGE: 0.0  
BLEU 2:  $3.55 \times 10^{-18}$  CIDEr: 0.0  
BLEU 3:  $1.12 \times 10^{-16}$  CLIP-S: 0.6218  
BLEU 4:  $4.51 \times 10^{-16}$  METEOR: 0.0  
METEOR: 0.0 RefCLIP-S: 0.5998



**GT:** plants and herbs: marjoram, hill, potted plants, container of ceramics: jar, jug, pot, vase, pig, scholastic education, tuition, Contrariety; ‘Contrarietà’, Vulgarity, proverbs, sayings

**Caption:** ‘hoofed animals : boar, container of metal : bucket, can, canister, drum, tin, proverbs, sayings.

BLEU 1: 0.1425 METEOR: 0.079  
BLEU 2: 0.1049 ROUGE: 0.211  
BLEU 3:  $7.7407 \times 10^{-7}$  CIDEr: 0.057  
BLEU 4:  $2.153 \times 10^{-9}$  CLIP-S: 0.884  
METEOR: 0.079 RefCLIP-S: 0.817



**GT:** apostle, unspecified, key.

**Caption:** head turned to the right, historical persons.

BLEU 1:  $1.43 \times 10^{-16}$  ROUGE: 0.0  
BLEU 2:  $1.54 \times 10^{-16}$  CIDEr: 0.0  
BLEU 3:  $1.68 \times 10^{-16}$  CLIP-S: 0.5221  
BLEU 4:  $1.85 \times 10^{-16}$  METEOR: 0.0  
METEOR: 0.0 RefCLIP-S: 0.6427

**Figure 3.** Examples of images from the Iconclass Caption test set, their corresponding ground-truth and generated captions and the values of evaluation metrics for those examples.

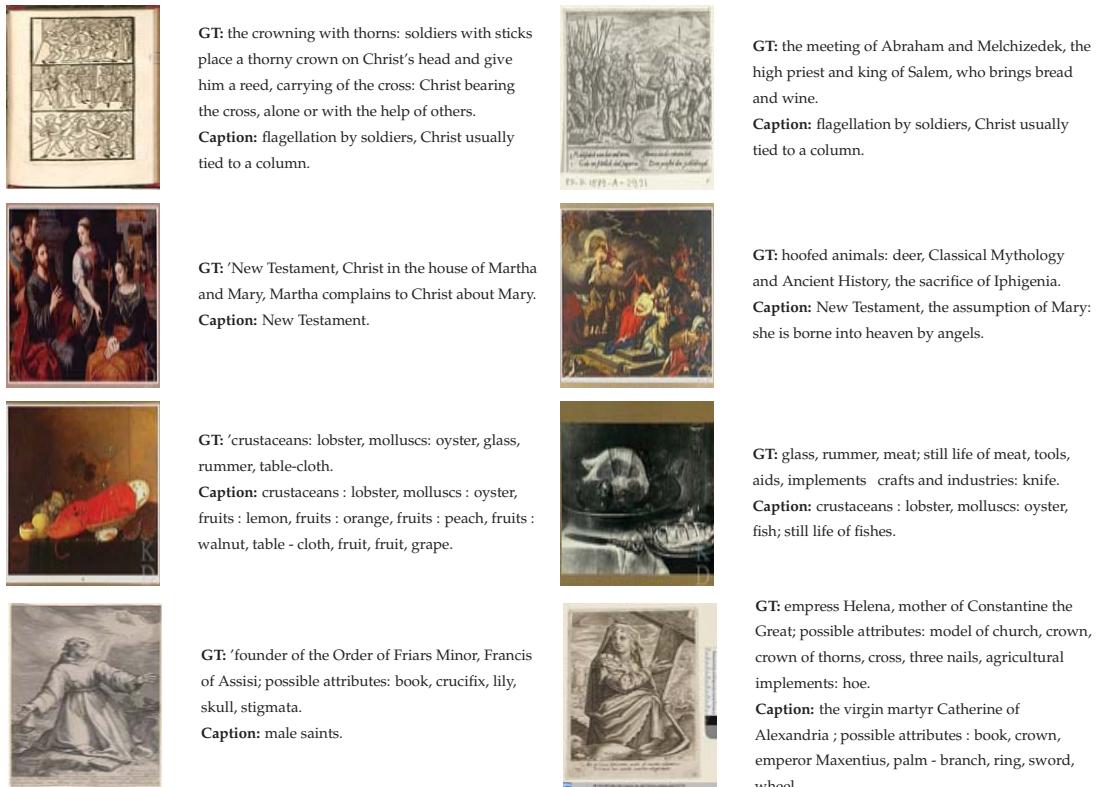
#### 4.2. Qualitative Analysis

Qualitative analysis was performed by exploring examples of images and generated captions on two datasets. One is the test set of the Iconclass Caption dataset that serves for direct comparison between the generated captions and ground-truth descriptions. The other dataset is a subset of the WikiArt painting collection, which does not include textual descriptions of images but has a broad set of labels associated with each image. Therefore, this dataset is useful to explore how the generated captions relate to the genre categorization of the paintings.

##### 4.2.1. Iconclass Caption Test Set

To gain a better insight into the generated image captions, in Figure 4 several examples are shown. The presented image-text pairs were chosen to demonstrate both good examples (the left column) and bad examples (the right column) of generated captions.

Analysis of the unsuccessful examples indicates that similarities between visual representations can result in generating analogous, but very misleading, iconographic captions. It also demonstrates underlying biases within the dataset. For instance, in the Iconclass Caption training test, there are more than a thousand examples that include the phrase “New Testament” in the description. Therefore, images that include structurally similar scenes, particularly from classical history and mythology, are sometimes wrongly attributed as depicting a scene from the New Testament. This signifies the importance of balanced examples in the training dataset and indicates directions for possible future improvements. Furthermore, by analysing various examples of generated captions, it becomes clear that recognizing fine-grained categories, e.g., exact names of saints or specific historical scenes, is still a very challenging task.

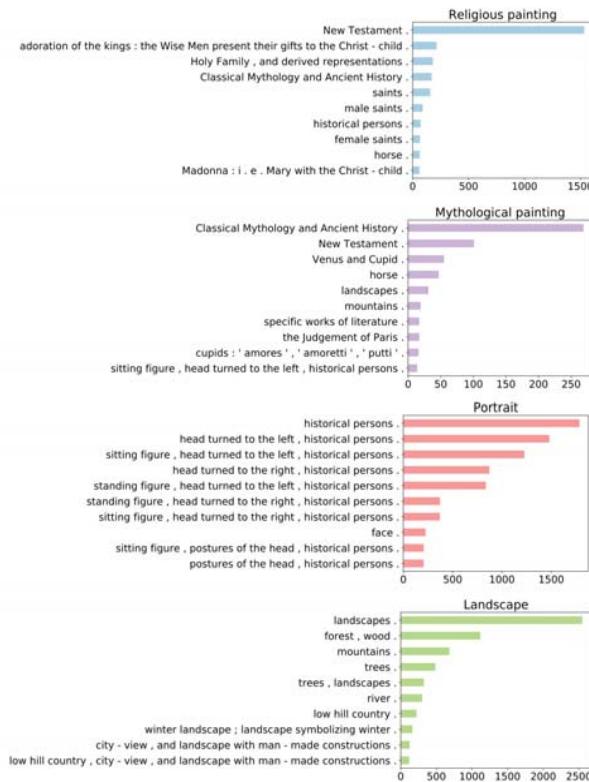


**Figure 4.** Examples of images from the Iconclass Caption test set, with their corresponding ground-truth and generated captions. Examples shown on the left side represent cases where the generated captions are successfully aligned with the iconographic content of the image, while examples shown on the right demonstrate unsuccessful examples.

The Iconclass dataset is a collection of very diverse images and apart from the Iconclass classification codes, there are currently no other metadata available for the images. Therefore, it is difficult to perform an in-depth exploratory analysis of the dataset and the generated results in regard to attributes relevant in the context of art history such as the date of creation, style, genre, etc. For this reason, the fine-tuned image captioning model was employed on another artwork dataset.

#### 4.2.2. WikiArt Dataset

The quality of the generated captions and the model's capacity to generalize to new data are further explored by employing the model on another artwork dataset, a subset of the WikiArt dataset that includes labels related to the genre of the paintings. Figure 5 shows the distribution of the most commonly generated descriptions in relation to four different genre categories. From this basic analysis, it is obvious that the generated captions are meaningful in relation to the content and the genre categorization of images.



**Figure 5.** Distribution of most commonly generated captions in relation to four different genres in the WikiArt dataset.

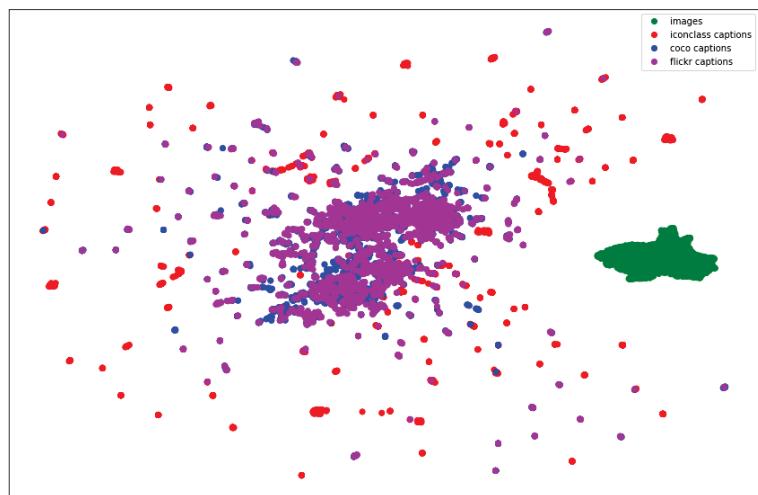
To understand the contribution of the proposed model in the context of iconographic image captioning, it is interesting to compare the Iconclass captions with captions obtained from models trained on natural images. For this purpose, two models of the same architecture but fine-tuned on the Flickr 30 i MS COCO datasets were used. Figure 6 shows several examples from the WikiArt dataset with corresponding Iconclass, Flickr and COCO captions. It is evident that the other two models generate results that are meaningful in relation to the image content but do not necessarily contribute to producing more fine-grained and context-aware descriptions. However, the values of the CLIP-Score evaluation metric are, in general, higher for captions generated using the model pretrained on natural images than the Iconclass model.

The mean value of CLIP-S on the Iconclass captions of the WikiArt subset is 0.595, while the mean score of the Flickr caption is 0.684 and that of the Coco captions is 0.691. This result corresponds to the conclusion presented in [10], which suggests that, when assessing a direct description and a more non-literal caption, the CLIPScore will generally prefer the literal one. However, because the CLIP model is trained on a very large set of examples extracted from the internet, it has probably encountered some well-known cases of iconographic image-text relations in the training set. This explains the high values of the CLIPScore for the third and fourth examples in Figure 6.

To gain a better understanding of the CLIPScore in relation to the various types of image captions and the images themselves, Figure 7 shows a projection of the image and different caption features obtained using the CLIP ViT-B/32 model.

	<i>Anthony van Dyck, Venus asking Vulcan for the Armour of Aeneas, c.1632</i> Iconclass caption: Classical Mythology and Ancient History. Flickr caption: A painting of a group of people. Coco caption: A painting of a group of people in a field.	Iconclass CLIP-S: 0.672 Flickr CLIP-S: 0.640 Coco CLIP-S: 0.539
	<i>Hans Memling, Man of Sorrows, c.1490</i> Iconclass caption: Christ. Flickr caption: A marble statue of a seated man. Coco caption: A painting of a man holding a hammer.	Iconclass CLIP-S: 0.602 Flickr CLIP-S: 0.565 Coco CLIP-S: 0.659
	<i>Lucas Cranach the Elder, Fall of Man, 1537</i> Iconclass caption: Eve offers the fruit to Adam. Flickr caption: Two young boys are climbing a tree. Coco caption: A statue of a boy and a girl near a tree.	Iconclass CLIP-S: 0.758 Flickr CLIP-S: 0.698 Coco CLIP-S: 0.622
	<i>Antoine Watteau, Cupid Disarmed, c.1715</i> Iconclass caption: Venus and Cupid . Flickr caption: 3 children in a circle . Coco caption: A portrait of a woman holding a child.	Iconclass CLIP-S: 0.799 Flickr CLIP-S: 0.595 Coco CLIP-S: 0.681

**Figure 6.** Examples from the WikiArt dataset with captions generated by models fine-tuned on the Iconclass, Flickr and COCO datasets.



**Figure 7.** UMAP (Uniform Manifold Approximation and Projection) plot depicting the CLIP (Contrastive Language-Image Pre-Training) model embeddings of the images and various generated captions on a subset of the WikiArt dataset.

The distribution of data points in Figure 7 indicates that the captions generated using the COCO and Flickr fine-tuned models are more aligned with each other, while the Iconclass captions are more dispersed. This is understandable considering the difference in the vocabulary and structure of the Iconclass descriptions. Overall, although the CLIPScore

shows very good results in assessing the similarity of the image content and textual description, as well as particularly promising results in recognizing iconographic relations, it is still necessary to achieve a higher level of explainability of the CLIP model in order to determine its applicability for evaluating iconographic captions.

## 5. Conclusions

This paper introduces a novel model for generating iconographic image captions. This is achieved by utilizing a large-scale dataset of artwork images annotated with concepts from the Iconclass classification system designed for art and iconography. Within the scope of this work, the available annotations were processed into clean textual descriptions and the existing dataset was transformed into a collection of suitable image-text pairs. The dataset was used to fine-tune a transformer-based vision-language model. For this purpose, object classification aware region features were extracted from the images using the Faster R-CNN model. The base model in our fine-tuning experiment is an existing model, called the VLP model, that was pretrained on a natural image dataset on intermediate tasks with unsupervised learning objectives. Fine-tuning pretrained vision-language models represents the current state-of-the-art approach for many different multimodal tasks.

The captions generated by the fine-tuned models were evaluated using standard image captioning metrics and recently introduced reference-free metrics. Due to the specific properties of the Iconclass dataset, standard image captioning evaluation metrics are not very informative regarding the relevance and appropriateness of the generated captions in relation to the image content. The reference-free metric, CLIPScore, represents an interesting new approach for evaluating image captions based on the cosine distance between image and text embeddings from a joint feature space. This image captioning metric shows very promising results in evaluating the semantic relation of images and texts, particularly in the case of well-known iconographic image-text examples. However, it is still uncertain if it the best choice for assessing all iconographic image captions because it generally favours literal over non-literal image-text relations. In this context, one of the major directions for future research is related to exploring multimodal deep learning approaches in the context of non-literal relations between images and texts.

The overall quantitative and qualitative evaluations of the results suggest that it is possible to generate iconographically meaningful captions that capture not only the depicted objects but also the art historical context and relation between subjects. However, there is still room for significant improvement. In particular, the unbalanced distribution of themes and topics within the training set results in often wrongly identified subjects in the generated image descriptions. Furthermore, the generated textual descriptions are often very short and could serve more as labels rather than captions. Nevertheless, the current results show significant improvement in comparison to captions generated from artwork images using models trained on natural image caption datasets. Further improvement can potentially be achieved with fine-tuning the current model on a smaller dataset with more elaborate ground-truth iconographic captions.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Additional material including the model weights and dataset are available at <https://github.com/EvaCet/Iconclass-image-captioning>.

**Conflicts of Interest:** The author declares no conflict of interest.

## References

- Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014, Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Part V, Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2014; Volume 8693, pp. 740–755.
- Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [[CrossRef](#)]
- Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [[CrossRef](#)]
- Panofsky, E. *Studies in Iconology. Humanistic Themes in the Art of the Renaissance*, New York; Harper and Row: New York, NY, USA, 1972.
- Posthumus, E. Brill Iconclass AI Test Set. 2020. Available online: <https://labs.brill.com/ictestset/> (accessed on 20 July 2021).
- Couplie, L.D. Iconclass: An iconographic classification system. *Art Libr. J.* **1983**, *8*, 32–49. [[CrossRef](#)]
- Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.J.; Gao, J. Unified Vision-Language Pre-Training for Image Captioning and VQA. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; AAAI: Menlo Park, CA, USA, Volume 34, No. 07, pp. 13041–13049.
- Cetinic, E. Iconographic Image Captioning for Artworks. In Proceedings of the ICPR International Workshops and Challenges, Virtual Event, Milan, Italy, 10–15 January 2021; Springer: New York, NY, USA, pp. 502–516.
- Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. *arXiv* **2021**, arXiv:abs/2103.00020.
- Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R.L.; Choi, Y. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *arXiv* **2021**, arXiv:2104.08718.
- Cetinic, E.; Lipic, T.; Grgic, S. Fine-tuning convolutional neural networks for fine art classification. *Expert Syst. Appl.* **2018**, *114*, 107–118. [[CrossRef](#)]
- Sandoval, C.; Pirogova, E.; Lech, M. Two-stage deep learning approach to the classification of fine-art paintings. *IEEE Access* **2019**, *7*, 41770–41781. [[CrossRef](#)]
- Milani, F.; Fraternali, P. A Data Set and a Convolutional Model for Iconography Classification in Paintings. *arXiv* **2020**, arXiv:2010.11697.
- Seguin, B.; Striolo, C.; Kaplan, F. Visual link retrieval in a database of paintings. In Proceedings of the Computer Vision (ECCV) 2016, Amsterdam, The Netherlands, 8–16 October 2016; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9913, pp. 753–767.
- Mao, H.; Cheung, M.; She, J. Deepart: Learning joint representations of visual arts. In Proceedings of the 25th ACM international conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1183–1191.
- Castellano, G.; Vessio, G. Towards a tool for visual link retrieval and knowledge discovery in painting datasets. In *Digital Libraries: The Era of Big Data and Data Science, Proceedings of the 16th Italian Research Conference on Digital Libraries (IRCDL) 2020, Bari, Italy, 30–31 January 2020*; Springer: Berlin, Germany, 2020; Volume 1177, pp. 105–110.
- Crowley, E.J.; Zisserman, A. In search of art. In Proceedings of the Computer Vision (ECCV) 2014 Workshops, Zurich, Switzerland, 6–12 September 2014; Lecture Notes in Computer Science; Springer: Berlin, Germany, Volume 8925, pp. 54–70.
- Strezoski, G.; Worring, M. Omniart: A large-scale artistic benchmark. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2018**, *14*, 1–21. [[CrossRef](#)]
- Madhu, P.; Kosti, R.; Mühlberg, L.; Bell, P.; Maier, A.; Christlein, V. Recognizing Characters in Art History Using Deep Learning. In Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia heritAge Contents, Nice, France, 21–25 October 2019; pp. 15–22.
- Jenicek, T.; Chum, O. Linking Art through Human Poses. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; pp. 1338–1345.
- Shen, X.; Efros, A.A.; Aubry, M. Discovering visual patterns in art collections with spatially-consistent feature learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9278–9287.
- Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Huang, F.; Deussen, O.; Xu, C. Exploring the Representativity of Art Paintings. *IEEE Trans. Multimed.* **2020**. [[CrossRef](#)]
- Cetinic, E.; Lipic, T.; Grgic, S. Learning the Principles of Art History with convolutional neural networks. *Pattern Recognit. Lett.* **2020**, *129*, 56–62. [[CrossRef](#)]
- Elgammal, A.; Liu, B.; Kim, D.; Elhoseiny, M.; Mazzone, M. The shape of art history in the eyes of the machine. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, New Orleans, LA, USA, 2–7 February 2018; AAAI Press: Palo Alto, CA, USA, 2018; pp. 2183–2191.
- Hayn-Leichsenring, G.U.; Lehmann, T.; Redies, C. Subjective ratings of beauty and aesthetics: Correlations with statistical image properties in western oil paintings. *i-Perception* **2017**, *8*, 2041669517715474. [[CrossRef](#)]
- Cetinic, E.; Lipic, T.; Grgic, S. A deep learning perspective on beauty, sentiment, and remembrance of art. *IEEE Access* **2019**, *7*, 73694–73710. [[CrossRef](#)]

27. Sargentis, G.; Dimitriadis, P.; Koutsoyiannis, D. Aesthetical Issues of Leonardo Da Vinci's and Pablo Picasso's Paintings with Stochastic Evaluation. *Heritage* **2020**, *3*, 283–305. [[CrossRef](#)]
28. Cetinic, E.; She, J. Understanding and Creating Art with AI: Review and Outlook. *arXiv* **2021**, arXiv:2102.09109.
29. Castellano, G.; Vessio, G. Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. *Neural Comput. Appl.* **2021**, *1*–20. [[CrossRef](#)]
30. Fontanella, F.; Colace, F.; Molinara, M.; Di Freca, A.S.; Stanco, F. Pattern Recognition and Artificial Intelligence Techniques for Cultural Heritage. *Pattern Recognit. Lett.* **2020**, *138*, 23–29. [[CrossRef](#)]
31. Garcia, N.; Vogiatzis, G. How to read paintings: Semantic art understanding with multi-modal retrieval. In Proceedings of the European Conference on Computer Vision (ECCV) 2018 Workshops, Munich, Germany, 8–14 September 2018; Lecture Notes in Computer Science; Springer: Berlin, Germany, Volume 11130, pp 676–691.
32. Baraldi, L.; Cornia, M.; Grana, C.; Cucchiara, R. Aligning text and document illustrations: Towards visually explainable digital humanities. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 1097–1102.
33. Stefanini, M.; Cornia, M.; Baraldi, L.; Corsini, M.; Cucchiara, R. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In Proceedings of the Image Analysis and Processing (ICIAP) 2019, 20th International Conference, Trento, Italy, 9–13 September 2019; Lecture Notes in Computer Science; Springer: Berlin, Germany, Volume 11752, pp. 729–740.
34. Cornia, M.; Stefanini, M.; Baraldi, L.; Corsini, M.; Cucchiara, R. Explaining digital humanities by aligning images and textual descriptions. *Pattern Recognit. Lett.* **2020**, *129*, 166–172. [[CrossRef](#)]
35. Banar, N.; Daelemans, W.; Kestemont, M. Multi-modal Label Retrieval for the Visual Arts: The Case of Iconclass. In Proceedings of the 13th International Conference on Agents and Artificial Intelligence, (ICAART) 2021, Online Streaming, 4–6 February 2021; SciTePress: Setúbal, Portugal, Volume 1, pp. 622–629.
36. Bongini, P.; Becattini, F.; Bagdanov, A.D.; Del Bimbo, A. Visual Question Answering for Cultural Heritage. *arXiv* **2020**, arXiv:2003.09853.
37. Garcia, N.; Ye, C.; Liu, Z.; Hu, Q.; Otani, M.; Chu, C.; Nakashima, Y.; Mitamura, T. A Dataset and Baselines for Visual Question Answering on Art. *arXiv* **2020**, arXiv:2008.12520.
38. Sheng, S.; Moens, M.F. Generating Captions for Images of Ancient Artworks. In Proceedings of the 27th ACM International Conference on Multimedia, (MM) 2019, Nice, France, 21–25 October 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 2478–2486.
39. Gupta, J.; Madhu, P.; Kosti, R.; Bell, P.; Maier, A.; Christlein, V. Towards Image Caption Generation for Art Historical Data. In Proceedings of the AI Methods for Digital Heritage, Workshop at KI2020 43rd German Conference on Artificial Intelligence, Bamberg, Germany, 21–25 September 2020.
40. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
41. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
42. Tan, H.; Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv* **2019**, arXiv:1908.07490.
43. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 13–23.
44. Chen, Y.C.; Li, L.; Yu, L.; Kholy, A.E.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. Uniter: Learning universal image-text representations. *arXiv* **2019**, arXiv:1909.11740.
45. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
46. Sharma, P.; Ding, N.; Goodman, S.; Sorice, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2556–2565.
47. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
48. Denkowski, M.; Lavie, A. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA, 26–27 June 2014; pp. 376–380.
49. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
50. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
51. Xia, Q.; Huang, H.; Duan, N.; Zhang, D.; Ji, L.; Sui, Z.; Cui, E.; Bharti, T.; Zhou, M. Xgpt: Cross-modal generative pre-training for image captioning. *arXiv* **2020**, arXiv:2003.01473.



## Article

# A Methodology for Semantic Enrichment of Cultural Heritage Images Using Artificial Intelligence Technologies

Yalemisew Abgaz <sup>1,\*</sup>, Renato Rocha Souza <sup>2</sup>, Japesh Methuku <sup>1</sup>, Gerda Koch <sup>3</sup> and Amelie Dorn <sup>2</sup>

<sup>1</sup> ADAPT Centre, School of Computing, Dublin City University, Glasnevin Campus, Dublin 9, Dublin, Ireland; Japesh.Methuku@adaptcentre.ie

<sup>2</sup> Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH OeAW), Austrian Academy of Sciences, 1010 Vienna, Austria; renato.souza@oeaw.ac.at (R.R.S.); Amelie.Dorn@oeaw.ac.at (A.D.)

<sup>3</sup> AIT Angewandte Informationstechnik Forschungsgesellschaft mbH, Europeana Local AT, Klosterwiesgasse 32, 8010 Graz, Austria; Kochg@europeana-local.at

\* Correspondence: Yalemisewm.Abgaz@dcu.ie



**Citation:** Abgaz, Y.; Rocha Souza, R.; Methuku, J.; Koch, G.; Dorn, A. A Methodology for Semantic Enrichment of Cultural Heritage Images Using Artificial Intelligence Technologies. *J. Imaging* **2021**, *7*, 121. <https://doi.org/10.3390/jimaging7080121>

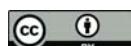
Academic Editor: Giovanna Castellano, Gennaro Vessio and Fabio Bellavia

Received: 15 June 2021

Accepted: 18 July 2021

Published: 22 July 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The digitisation of cultural heritage resources has opened a new way of sharing and utilising information that was previously offline. Many Galleries, Libraries, Archives, and Museums (GLAMs) transform tangible and intangible heritage by converting the physical resources into digital images, audios, videos, simulation models, and virtual realities [1]. As part of the effort, cultural images that represent the culture and history of societies become available in digital formats on the semantic web [2–4]. UNESCO defines cultural heritage to encompass tangible heritage including movable (paintings, sculptures, coins, and manuscripts), immovable (monuments, archaeological sites), and underwater cultural heritage; and intangible heritage covering oral traditions, performing arts, and rituals (<http://www.unesco.org/new/en/culture/themes/illicit-trafficking-of-cultural-property/unesco-database-of-national-cultural-heritage-laws/frequently-asked-questions/definition-of-the-cultural-heritage/>) (accessed on 15 April 2021)). Cultural heritage images include paintings, photographs, drawings, and sketches that represent the culture and/or history of a particular society or country [4]. Although cultural images are available embedded on a physical medium, the massive digitisation process makes them accessible in a digital intangible format. In this paper, digital cultural heritage images (we refer to them as cultural images now onwards) represent a selection of digital images that reflect the

culture of a society in the past and present. Since culture encompasses a wider range of human aspects, it is difficult to fully understand and cover all these aspects. This paper focuses on cultural images that are related to edible food.

Despite the growing number of cultural images, the availability and the maturity of methods and tools to systematically exploit the content of the images in a structured and meaningful way is still at its early stage [5,6]. Solutions that work well in other domains (such as medical imaging) were not exploited until recently. Elsewhere, digital images are widely represented by metadata related to the creators, creation time, title, and short descriptions of the images. However, these representations lack contextual information about the cultural content of the images. For this reason, the most valuable information embedded in the images is not explicitly annotated and utilised.

In the light of recent advancements in AI, there are now greater opportunities for digital humanities to apply sophisticated AI solutions to enrich and exploit cultural images [7]. Natural language processing [8], image classification [9,10], Computer Vision (CV) [11], and Virtual Reality (VR) are some of the areas that are gaining strong momentum and fast adoption in digital humanities research. CV in particular has been used to analyse cultural heritage collections including architectures, buildings, and other cultural artefacts. The analysis includes object detection, classification, reconstruction, and semantic annotation. Ontologies [12] have been proven to be crucial for semantic enrichment of cultural images. Ontologies are used to consistently represent resources to be understood and interpreted uniformly by humans and machines [13] in the Linked Open Data (LOD) space [14].

Despite the availability of technological solutions, the digital humanities domain has not yet exploited the full benefits due to the lack of an end-to-end methodology that supports the transformation of cultural images from mere digital entities to useful resources for supporting scientific research. Unless addressed methodologically, the use of existing technologies for searching, analysing, and annotating cultural images with such usable content can be breathtakingly time-consuming. Moreover, the absence of ground truth which would normally serve as a basis for the development and evaluation of AI solutions is lacking. An interpretation template for both concrete and abstract concepts of cultural images is missing [15]. Thus, a combination of manual and automatic annotation is widely proposed to semantically enrich cultural images.

To date, CV and image analysis technologies focus on detecting concrete objects in the images [15]. Although the technologies serve well for object detection, they are in their early stage generating usable annotations for abstract and highly subjective aspects [16] of cultural images. Moreover, most of the CV technologies are trained using images that emerged from domains that have sufficient and quality data for the training and evaluation of such systems [17].

This paper presents a three-phase methodology for semantic enrichment of cultural images using AI technologies. Our methodology begins with the preparation phase which enables us to understand the domain, acquire the target resources including the images, ontologies, and vocabularies. The second phase presents tools and techniques for analysing and annotating the content, training and evaluating the CV models, whereas phase three focuses on deployment, exploration, and integration of active learning components in the system. Our methodology follows a continuous and iterative development within and across the phases. The proposed methodology is developed in the framework of the ChIA project (ChIA—accessing and analysing cultural images with new technologies) [18] where AI solutions are applied to solve problems in the digital humanities domain. The contribution of the paper includes:

- An end-to-end methodology and case study for semantic enrichment of cultural images.
- A technique for building and exploiting CV tools for digital humanities by employing iterative annotation of sample images by experts.
- A vocabulary for enriching cultural images in general and images related to food and drink in particular.
- A benchmarking data set which could serve as a ground truth for future research.

- A discussion of the lessons, challenges, and future directions.

The remaining sections of the paper are organised as follows. Section 2 introduces the complex aspects of cultural images and how they are represented and analysed using CV and semantic web technologies. Section 3 outlines our proposed methodology which is organised into three phases: preparation, analysis, and integration and exploitation. In Section 4, we present a case study where our proposed methodology is applied to cultural images collected from Europeana and Europeana-local Austria. Following the findings of the case study, the discussion of the results is presented in Section 5. Finally, recommendations and future work are presented in Section 6.

## 2. State of the Art

Until recently, the focus of digital humanities research was on the conversion of resources into a digital representation and publishing mainly in platforms supported by the respective institutions [7,19]. However, the emergence of new image processing technologies, deep learning, natural language processing, and semantic web technologies provides new opportunities to enhance the organisation, interlinking and exploitation of cultural heritage images. In this section, we summarise the advancements in these areas.

### 2.1. Computer Vision in Digital Humanities

Computer Vision applications such as image recognition, object detection, and classification using large-scale digital images have gained significant traction in digital humanities research. In this section, we review the advancements in Convolutional Neural Network (CNN) in light of their application in the digital humanities domain.

#### Convolutional Neural Networks

CNN comprises different layers like convolution, pooling, and activation that help in analyzing the patterns in an image. The convolution layers form a core building block of a CNN where each layer consists of a set of K learnable filters, each filter having a width and height. The output of each convolution operation produces an activation map which is a 2-dimensional output. The images are represented by pixels and mathematical operations are used by CNN for analyzing the patterns embedded within the images. CNNs are built using a sequence of convolution, pooling and non-linearity layers where convolution is used to extract spatial features and pooling layers are used to reduce the spatial dimensions of the image.

ImageNet is a benchmark data set having around 15 million labelled images that represent 22,000 categories. ImageNet Large Scale Visual Recognition Challenge (ILSVRC) uses around 1.2 million images for training, 50,000 images for validation and 100,000 images for testing. CNN architectures are designed to classify the images for ILSVRC and the architectures have evolved. LeNet5 [20] was one of the simplest CNN architectures having two convolution and three fully connected layers. The architecture in which the convolution and pooling layers were stacked in LeNet5 turned a baseline for other CNN models. AlexNet [21] was the next benchmark CNN architecture that was a much deeper and wider version of LeNet5 and could learn much more complex objects and used Rectified Linear Units (ReLU) as non-linearities. The architecture also saw the use of dropout regularisation which is a technique in deep learning to reduce the effect of overfitting (models' ability to generalize on unseen images is suppressed) and also data augmentation techniques which allows the CNN model to visualize the images by applying different properties like translation, reflections, and patch extractions. The data augmentation technique is particularly useful when there is minimum availability of images for training a CNN model as it introduces new variations into the data set. AlexNet has eight layers with five convolutional and three fully connected layers.

What changed between LeNet5 and AlexNet is the number of layers stacked to design a CNN architecture. With the increase in depth of the layers in a CNN, there was an improved chance of learning complex patterns and representations, and these

patterns resulted in more complex architectures going much deeper and with more trainable parameters. VGG (Visual Geometry Group) network [22] was designed and developed by the researchers at Oxford University which has thirteen convolutional and three fully-connected layers with ReLU as non-linearity. There are two variants of the VGG network, VGG16, and VGG19 and use smaller  $3 \times 3$  filters in each convolutional filter. These multiple smaller filters can emulate the effect of larger receptive fields to represent complex features. However, a network with such large depth also makes the model bigger, and VGG network has 138 million trainable parameters.

ResNet50 [23] was trained on ImageNet data set with a 152 layer deep convolutional neural network, which is eight times deeper than the VGG network. An ensemble of the residual networks achieved a 3.57% error on the ImageNet test set. The experiments were conducted to understand the use of residual learning and shortcut connections for improving the generalizability of the model. Convolution and identity blocks form the basic building block of ResNet50 and this CNN model has 26 M parameters.

Inception\_V3 [24] is a variant in the inception family of pre-trained convolutional neural networks, the architecture of which is reviewed by rethinking the inception architecture to realize computational efficiency and fewer parameters. The Inception\_V3 architecture is composed of factorized convolutions where the aim is to reduce the number of connections/parameters without decreasing the performance/efficiency of the neural network. The idea behind factorized convolutions is to replace a convolution of a larger receptive field with smaller size convolutions. For example, one  $5 \times 5$  convolution layer can be decomposed into two  $3 \times 3$  convolution layers, which further reduces the number of parameters. Furthermore, a kind of dropout regularization technique, label smoothing is used to prevent the logits from taking large values. Label smoothing also helps in preventing the CNN model from overfitting.

With the evolution of CNN architectures, there has been a lot of research to reduce the complexity of the model by making the models much deeper. In total,  $1 \times 1$  (pointwise) convolutions were adapted in the models using which the features across the feature maps could be spatially combined with the effective use of very few parameters. Depthwise convolutions is one such idea that comprises two convolution operations, spatial convolution followed by pointwise convolutions. This made the CNN networks lighter and faster due to fewer trainable parameters and fewer FLOPs (floating-point operations). Xception [25] is an improvement and an extension of the inception family of CNN architectures with few architectural changes and effective as ResNet50 and Inception\_V3. In Xception, depthwise separable convolutions have replaced the inception modules. There is a performance improvement in Xception due to the more efficient use of model parameters. The pointwise convolutions are followed by depthwise convolutions, unlike the inception network. The Xception architecture is divided into three flows, entry flow, middle flow, and exit flow. The data is passed through the entry flow, and the middle flow is repeated eight times and then the data is passed through the exit flow.

Machine Learning has been used in the space of digital humanities to classify images belonging to cultural heritage in [10], where there is a comparison of different approaches like multilayer perceptron, k-nearest neighbour, and CNNs. The classification was based on concrete concepts that are well defined and the patterns within the image attribute to a particular class/category. We aim to investigate and analyze how CNNs can be used for abstract concepts (Sections 3 and 4). We have used three pre-trained models, Inception\_V3, ResNet50, and Xception to classify the images based on abstract concepts.

There are a few shortcomings of computer vision and deep learning algorithms used for the classification of images. First, they are mostly trained to support general-purpose applications which might not be effective for very specific domains. Second, the models are trained to recognise well-known concrete objects, shapes, colours, etc. However, in the cultural images, the interest encompasses the identification of abstract concepts represented in the cultural images such as formality, appealingness, etc. Another shortcoming is that the techniques used in designing these models work mathematically well, but are often

claimed as being black boxes where there are no set of rules for maximizing the results. This is deeply concerning because it minimizes the opportunity to verify the decision-making process while working towards the objectives.

## 2.2. Semantic Web Technologies

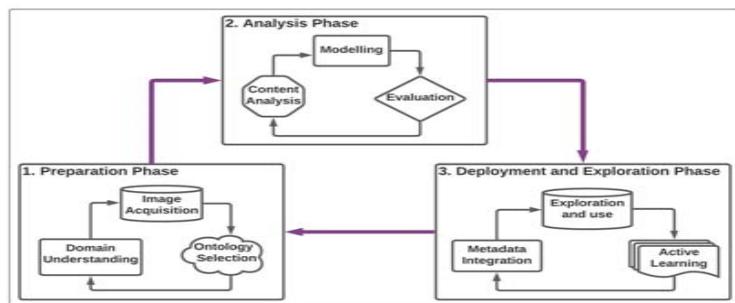
The semantic web refers to an extension of the World Wide Web with a goal of encoding semantics to the data on the web to facilitate the interlinking of web resources and to support machine-readable format. The semantic web uses technologies such as Resource Description Framework (RDF (<https://www.w3.org/TR/owl-features/>) (accessed on 13 July 2021))) and Web Ontology Language (OWL (<https://www.w3.org/TR/rdf-syntax-grammar/>) (accessed on 13 July 2021))) to facilitate encoding and processing meaning for the consumption of human and computer agents [26]. Semantic web technologies benefited from the development of large repositories (DBpedia [27], Europeana, and swissbib (<https://data.swissbib.ch/>) (accessed on 10 May 2021))), multilingual and interdisciplinary vocabularies (BabelNet [28], WordNet [29]), specialised ontologies (CIDOC-CRM [30], EDM [31]), and the LOD initiatives. Such repositories not only provide the required vocabularies to enrich cultural images but also enable semantic interlinking of the resources and creating links that can be exploited by both human and artificial intelligent agents.

Previous research exploits the semantic web technologies in different forms. An ontology model for narrative image annotation has been developed to annotate images in the field of cultural heritage [14]. The authors developed an ontology model and a tool to semantically annotate narrative images. However, the image annotation is done manually being supported by the tool. Marcia [13] presented a review of semantic enrichment efforts in Libraries, Archives and Museums (LAM). The application of semantic enrichment in LAM includes the development of ontologies and semantic annotation of structured and unstructured digital resources. Although this is a review paper, it identified several semantic enrichment projects using ontologies, linked data and SPARQL queries to organise, search and retrieve digital resources. Another effort towards accessing historical and musical linked data is proposed in [32]. A web-based thin middleware that facilitates the use of SPARQL queries to access digital humanities linked data sets on the web is proposed. This paper presents a prototypical tool that allows the use of API-based access to enable users to interact with the linked data without using SPARQL queries. Although this paper focuses on the exploitation of semantic data sets, it also demonstrates the gap in the digital humanities domain.

Currently, the major challenge in this area includes the coverage of specialised ontologies that represent domain-specific concepts to interpret and understand consistently [33]. Most of the ontologies do not always cater to the needs of new applications. In this regard, although there is a continuous development of domain-specific ontologies and vocabularies representing major cultural aspects, it requires a substantial effort to integrate the ontologies/vocabularies to make a significant impact. Another observed gap in the literature is the slow adoption of the application of CV models to automatically detect and annotate abstract cultural aspects. CV models are capable of detecting objects and generating labels that can be fed with standard ontologies to generate labels represented by unique URIs to ensure consistent representation of the images.

## 3. Methodology for Enhancing the Visibility of Cultural Images

The proposed methodology (Refer to Figure 1) is organised into preparation, analysis, and integration and exploitation phases. The first phase focuses on acquiring, understanding, and representing the target domain and its related data. The second phase deals with the extraction of the content of the images using CV models and the last phase focuses on the integration and exploitation of the results of previous phases to provide rich information. The methodology follows an iterative and continuous improvement in each of the phases.



**Figure 1.** A three-phase-methodology for semantic enrichment of cultural images.

### 3.1. Phase-1: Preparation Phase

Some of the challenges faced in the semantic enrichment of cultural images include the complexity and diversity of the collection [34]. Most often, there is no one-fits-all solution that serves well all kinds of collections. Thus, a preparation phase that defines the domain of interest, acquiring representative data, and selecting the appropriate vocabulary is crucial to any digitisation and semantic enrichment project.

#### 3.1.1. Domain Understanding

Cultural images represent tangible artefacts such as buildings, food, cloth, machinery, and intangible artefacts such as festivities, language, music, and others [4]. Understanding the domain and defining the boundaries of the collection at the very early stage enables the selection and filtering of the target images and potential domain-specific ontologies. Given a large collection of digital images, applying semantic enrichment on the full collection in one step will result in a broader but shallow semantic annotation, whereas, focusing on a particular topic enables a deeper and rich semantic representation.

Thus, the first step in this process is understanding the collection and defining the topics that will be included in the semantic enrichment process. Focusing on the topic, where the target images deal with a particular subject such as food, drink, farming, and wedding. The additional dimension of the domain could include temporal information such as ancient, medieval, or modern eras or artefacts from specific seasons. The type of images including paintings, drawings, sketches, photographs could be used as additional criteria to defining the domain and set the boundaries.

Although several GLAMs focus on specific subjects, and times, aggregation platforms such as Europeana [34] expose very wide and diverse cultural images which pose additional challenges. In such situations, this particular step becomes very crucial for narrowing down the domain.

#### 3.1.2. Image Acquisition

This step involves the process of acquiring cultural images that are relevant to the selected domain in a digitised format. The image acquisition process could be specific to collections that are already available on existing platforms or new ones. This step becomes time-consuming particularly when narrow subject areas are selected. Even, with the support of efficient search and retrieval tools, current platforms often do not provide accurate and reliable results due to the quality of the associated metadata and the lack of rich semantics. This step further requires the allocation of significant manpower to spare on manual inspection and filtering. Image acquisition is done by domain experts in the topic area or using specialised tools that facilitate the selection of images relevant to the domain.

#### 3.1.3. Ontology Selection

Another crucial step in the preparation phase is the selection of suitable and rich semantics. Ontologies provide the semantic meaning and representation of concepts of a

domain [12]. Although generic ontologies representing widely applicable concepts can be used, the main focus of this step is the identification and selection/composition of ontologies that represent the concepts of interest of the selected domain both in its breadth and depth.

The selection of ontologies that are suitable for the semantic annotation of cultural images is often guided by the task at hand. There are several widely used criteria for selecting the right ontologies for specific tasks [35]. Once candidate ontologies are identified, often the decision would be selecting one or more of the identified ontologies or deciding to create a new ontology from scratch based on functional and non-functional requirements. Some of the functional requirements to determine the availability and suitability of ontologies include the coverage of the target concepts, the number of relationships captured and represented, and the expressiveness of the ontology. The non-functional requirements include a continuous maintenance and sustainability, availability for free re-use, compatibility with the standard (e.g., ISO 25964 norm) and its support for linked open data usage (similar to SKOS-Format).

### 3.2. Phase-2: Analysis Phase

This phase focuses on the automatic extraction of the content of the images. This analysis is not a trivial task, particularly identifying abstract and subtle concepts from cultural images is often difficult and subjective. However, we believe that a systematic approach that integrates expert input and active learning methods can ensure the extraction of concepts at least to the level of agreement observed between experts.

#### 3.2.1. Analysis of the Content of Images

A semantic analysis of the target images preferably by several domain experts not only provides a useful, and often an accurate representation of the content of the images, but also exhibits the level of agreement, detail, and difficulty that involve in the semantic enrichment process. Each image is analysed by experts and annotated using the selected concepts. The annotation process of concrete concepts (e.g., fruit, animal, vehicle, etc.) usually shows a higher inter-annotator agreement whereas annotation of abstract images exhibits lower (sometimes random) agreement. To avoid the subjectivity of the annotation, the analysis also includes the percentage of agreement exhibited between the annotators by including the statistics as a probability along with the annotated concepts for each of the images.

Processing the inter-annotator agreement between the annotators and understanding the nature and level of agreement in the annotation process of the selected domain provides crucial information in setting a benchmark for the envisioned automated annotation tool. Where humans display a higher level of agreement, automated systems are also expected to perform to the same level as human performance. Whereas, when the level of agreement between human annotators becomes low, the implication is that there is a huge subjectivity in the task which also requires to be captured in the automated systems. Due to the subjectivity involved in labelling cultural images with abstract concepts, instead of generating deterministic annotations, the area would benefit from fuzzy annotation [36] of images representing some level of uncertainty [37].

#### 3.2.2. Preparation of Training Data

For most of the tasks involving the preparation of training data for CV experiments, a large collection of training data is required. However, with the emergence of pre-trained models, this can be leveraged by reusing those pre-trained models in combination with a small set of new annotations focusing on particular features of the images. This leverages a significant portion of the task of collecting training data. However, for tasks that require domain-specific and in-depth analysis of the contents of the images, finding sufficient training data is still a major problem. What works for a general semantic annotation task does not often work for domain-specific annotations due to the requirement of domain

experts. Thus, the training data is often restricted to a few thousands of images. Methods to tackle the problem include exploiting available domain experts to train annotators to achieve a better understanding of the domain, engaging experts in a more creative approach, or relying on existing metadata and NLP tools to see if any pattern from the annotations of domain experts could be learned and generalised.

Existing CV models allow the use of previously trained models with a different set of images to be used to train new and unseen images and categories. Although this improves the learning rate of the algorithms, any successful CV tool still requires a large data set for training, validation, and testing.

### 3.2.3. Training and Selection of Best Performing Model

Recently, several computer-based image annotation methods became available. Among these CNN methods are gaining significant popularity [38,39]. The major considerations in selecting these CNN methods depend on their accuracy in generating results that are similar or superior to the accuracy achieved by human experts. In this phase, researchers train several models to select either the best performing one or ensemble two or more models to achieve higher performance. The selection is guided by the accuracy of the models during the training, validation, and testing phases.

The next step after the selection of the best-performing model is to use the selected model to label unseen images to obtain new annotations. These annotations will further include the predicted probability. The model uses a confidence level (0–100%) and this confidence level will be used to represent the confidence of the predicted annotation. The resulting data from the annotation will generate a list of annotation triples for each of the target images in a form of a CSV file (Example: Annotation.csv). A generic annotation format as a file is presented in Table 1.

**Table 1.** Expected annotation of cultural heritage images with confidence.

Image_Name	Label	Confidence
https://image1	Concept 1	85%
https://image1	Concept 2	100%
https://image1	Concept 3	40%
...	...	...

### 3.3. Phase-3: Integration and Exploitation Phase

One of the important factors in semantic enrichment is the integration of new semantic annotations into the existing semantic repositories. The integration process introduces new metadata to further describe the target resource. In systems that already have semantic repositories, the integration considers the legacy system and tries to integrate the new metadata in the legacy system without breaking the consistency and the validity of the data.

For resources that do not have a legacy semantic repository, the task involves creating the metadata repository, which further includes the selection, implementation, and deployment of a semantic repository. However, the selection and deployment of the repositories are out of the scope of this research paper. Further reading on the topic is available in [40–43].

#### 3.3.1. Integration of Results

The integration of large-scale semantic annotations generated by the annotation models involves the transformation of the generated annotation into a semantic representation. This process involves the use of subject-predicate-object triples where the subject represents the unique identifier (URL) of the image. The predicate represents the relationship between the subject and the object. The object is the predicted label that is generated by the model. Where there are several annotations available for a target image, an s-p-o triple will be generated for each of the annotations. A mapping of the annotation file into RDF

format is carried out by using R2RML mapping [44]. A typical R2RML mapping converts the input annotation into its equivalent RDF file using the following R2RML mapping ([https://github.com/yalemisewAbgaz/ChIA\\_Semantic\\_Annotation.git](https://github.com/yalemisewAbgaz/ChIA_Semantic_Annotation.git) (accessed on 15 July 2021)).

```
@prefix <list all your prefixes here>.

<#TripleMap1> a rr:TriplesMap ;
rr:logicalTable [rr:tableName "PREDICTIONS"];
rr:subjectMap[rr:template "https://www.europeana.eu/en/item/{IMAGE_NAME}";
rr:class edm:webResource; ];
rr:prediccateObjectMap[rr:predicate dc:description;
rr:prediccate rdfs:comment;
rr:objectMap[rr:column "LABEL"; ];];
rr:prediccateObjectMap[rr:predicate dc:description;
rr:prediccate rdfs:comment;
rr:objectMap[rr:column "LABEL_CONF"; ];];

<#TripleMap2> a rr:TriplesMap ;
[rr:sqlQuery """ Select * from PREDICTIONS where LABEL ='Appealing' """;
rr:subjectMap[rr:template "https://www.europeana.eu/en/item/{IMAGE_NAME}";
rr:prediccateObjectMap[rr:predicate dc:subject;
rr:objectMap[rr:template "http://purl.obolibrary.org/obo/MFOEM_000039"; ];].
```

An important aspect of the integration process involves the inclusion of certainty in the resulting semantic annotation. The area we are investigating involves a certain level of subjectivity. To represent the level of subjectivity in our semantic annotation, we add additional triples representing the annotation confidence as part of the description of the image, however, the representation of confidence/fuzzy knowledge needs to be addressed in the future.

Finally, the RDF data need to be integrated into the existing system. Although this is usually the task of the aggregators to decide on how to consume the annotation, our method is capable of generating the final data in a format specified by the user which includes RDF, TURTLE, NQUAD, or JSON-LD.

### 3.3.2. Supporting Efficient Exploitation

This step focuses on the exploitation of semantic annotation by supporting efficient aggregation and exploration of the data. There are different ways of achieving this. First, by providing users new exploration paths (SPARQL Query Templates) to query the collections using the newly added ontology concepts and relationships as used in [45,46]. Second, by supporting visualisation of the collection using the new annotation as a criterion for aggregating images as in [47]. Third, the use of interactive chatbots that are trained based on the annotated data to support queries that are based on precompiled templates. Although the first two options can be implemented directly on existing semantic repositories, the last option requires further development of a chatbot that is trained on the data set [48,49].

## 4. Case Study

This case study is conducted in the framework of ChIA project (<https://chia.acdh.oeaw.ac.at/> (accessed on 15 July 2021)) with a clear aim of engaging and testing new AI technologies against the background of a selected data set of food images for the benefit of accessing and analysing cultural data. The case study is applied following our proposed method (Section 3) which ensures the efficient representation of the data employing state-of-the-art semantic web technologies (ontologies and thesauri) and efficient analysis of the content using contemporary AI tools (CV). It presents a comprehensive methodology for answering how cultural knowledge of abstract food topics can be gained in a more structured and efficient method, and how this method is generalised to other areas in the digital humanities domain.

#### 4.1. Phase-1: Preparation Phase

In the preparation phase, we identified potential platforms that contain huge collections of cultural images. The general Europeana platform partnering with Europeana-local Austria, for accessing the images and the infrastructure, is selected. Europeana Local is a network portal for local and regional cultural and scientific data. The focus of the Europeana Local project is the coordination and integration of the heterogeneous data sets at both national and local level. Furthermore, for supporting our endeavour, we focused on technologies that are related to AI, particularly related to CV and semantic web technology. We also considered semantic technologies that we would require to use along the several stages of the semantic enrichment process.

##### 4.1.1. Understanding and Defining the Domain

The subject of the images is restricted to the topic of food and drink. Some of the rationales for selecting the topic of food includes, first, the availability of large collections of images related to edible food from the Europeana database from which our project draws a considerable quantity of cultural heritage images representing a great variety of cultural content holders (such as museums, archives, libraries, botanic gardens) across Europe. Second, the topic of food is a common topic, that all people can relate to. This includes the kind of food we consume, how it is produced, the fashionable food—these facets are all closely related to our political and economic history. Third, there is a huge diversity of cultural information represented by food images. Even if defining a clear boundary of the food topic is difficult, we restricted the topic to the production, preparation, presentation, and consumption of edible food.

Three concepts and their complements were selected ranging from very objective and concrete objects (fruit/non-fruit) to abstract (formal/informal) to very abstract and subjective (appealing/non-appealing) concepts. The definitions for the respective image labels were composed of definitions available in monolingual dictionaries and encyclopedias, according to the best fit for the overall theme of the project. In this step, we used a web-based image annotation tool (MakeSense.AI (<https://www.makesense.ai/>) (accessed on 1 April 2021))) which provided the environment suitable for the tasks at hand. MakeSense.AI is a simple, freely available and customisable tool that made it suitable for annotating images by multiple annotators.

##### 4.1.2. Image Acquisition

With the ChIA search platform established by Europeana-local Austria, we extracted food-related images (Refer Figure 2 for sample images) including paintings, photographs, and drawings. These images are extracted from the Europeana international platform that allows users to search images that contain food-related terminologies [50]. We collected more than 42,000 images in the first instance, grouped into several sub-folders representing the time, country, format, theme, etc. of the images. The search platform further allows the extraction of the digital images along with the associated metadata using RDF, XML or JSON-LD format generated for each image.

We further filtered images that are not related to food and drink. Since the initial selection of the sub-folders is based on food-related terminologies, the precision of the search was low and resulted in images that are not related to food. For example, the search apple resulted in several images of the Apple company and images related to Adam and Eve (due to their association with the apple tree). Generic food image detection tools were employed to further filter out images that are not related to food and drink [51]. For the final selection, a manual inspection of candidate files was conducted by Europeana local-At experts.



**Figure 2.** Sample food images selected from Europeana.

#### 4.1.3. Ontology Selection

To represent the concepts of food and drink efficiently, first we evaluated existing ontologies that are relevant to the topic of the research which focuses on ontologies related to food and drink as well as ontologies focusing on the cultural images. Finding ontologies that satisfy both requirements is difficult. Ontologies such as FoodOn (<https://github.com/FoodOntology/foodon>) (accessed on 10 May 2021)), AGROVOC (<http://www.fao.org/agrovoc/>) (accessed on 12 May 2021)) [52] food ontology and others represent the topic of food and drink but lack the cultural representations, whereas ontologies such as Iconclass (<http://www.iconclass.org/help/outline>) (accessed on 10 May 2021)) and Getty Art and Architecture Thesaurus (AAT) (<https://www.getty.edu/research/tools/vocabularies/aat/index.html>) (accessed on 10 May 2021)) represent the cultural aspect along with some concepts related to food. Since any one of these ontologies does not fully satisfy our requirements (see Section 3.1.3, we created a vocabulary that maps existing and well-established food and art vocabularies to create an integrated food vocabulary focusing on food in cultural and historical imagery.

The two cultural ontologies selected, Iconclass and the Getty AAT, are both widely used vocabularies for describing image content in the arts. Iconclass was developed in the early 1950s by Henri van de Waal, professor of art history at Leiden University. Today, the thesaurus is maintained by the RKD Rijksbureau voor Kunsthistorische Documentatie (Dutch Institute for Art History). Iconography is the art and science of recording themes that frequently appear in works of art [53] and Iconclass is an iconographic classification system that offers a hierarchically organised set of concepts to describe the content of visual resources in representational Western art (ancient mythology and Christian religious iconography) [54].

The Getty AAT was created in 1980 and is supported by the Getty Art History Information Program since 1983 [55]. It is a large thesaurus that is continuously updated and currently comprises about 71,000 records and about 400,400 terms, including synonyms and related terms, relevant to the field of art (December 2020). The terms, descriptions, and other information for generic concepts concern art, architecture, conservation, archaeology, and other cultural heritage [56].

Some ontologies for food exist, but most of them have been developed for specific applications related to food and lack cultural aspects. Targeted ontologies have been developed for agriculture, certain popular products such as pizza (<https://github.com/owlcs/pizza-ontology>) (accessed on 10 May 2021)) and wine (<https://www.w3.org/TR/owl-guide/wine.rdf> (accessed on 10 May 2021)), or in the context of culinary recipes, cooking, kitchen utensils, or nutrition. The FoodOn ontology was among the first attempts to build an ontology for broader applications. It includes nearly 30,000 terms about food and food-related human activities, such as agriculture, medicine, food safety control, shopping behaviour, and sustainable development [57].

In 2019, researchers created the FoodOntoMap [58] resource with the support of the Slovenian Research Agency programme and the H2020 project SAAM. FoodOntoMap consists of food concepts extracted from recipes, and thus foods that are edible for humans, and for each food concept, semantic tags from four food ontologies were assigned. The four ontologies used for matching were the Hansard corpus (<https://www.english-corpora.org>.

[org/hansard/](https://bioportal.bioontology.org/ontologies/OF/?p=summary) (accessed on 14 May 2021)), the FoodOn, OntoFood (<https://bioportal.bioontology.org/ontologies/OF/?p=summary> (accessed on 12 May 2021)) and SNOMED CT food (<https://confluence.ihtsdotools.org/display/DOCEG> (accessed on 14 May 2021)) ontologies. FoodOn is very comprehensive, and also provides semantics for food safety, food security, agricultural and animal husbandry practices associated with food production, culinary, nutritional and chemical ingredients and processes. As we only needed a selection of FoodOn concepts (human edible foods) for ChIA, FoodOntoMap offered us a perfect baseline for the ChIA vocabulary.

FoodOnToMap also provided us with an excellent base of matching concepts and we used this resource to update and expand with exact and related matches to the Iconclass and AAT ontology. Our goal was to add equivalence relationships between concepts that occur in the selected different ontologies and refer to the same entity in the world. The matching results from FoodOntoMap to AAT and Iconclass provided us with the first version of an integrated vocabulary of culture-related food terms with 1003 concepts, 1508 exact, and 1543 related matches from all processed food and art ontologies.

The resulting vocabulary is available at (<http://chia.ait.co.at/vocab/ChIA/index.php>, (accessed on 12 May 2021)) which provides details of food-related concepts and cultural concepts merged to represent cultural and historical food and drink-related concepts. Finally, the integrated ChIA food vocabulary was very well suited to search the Europeana corpus for food-related images and thus facilitated the repeated creation of training sets for data annotation.

#### 4.2. Phase-2

##### 4.2.1. Analysis of the Contents of the Images

Due to the diversity and richness of the format and contents of the images, we conducted this phase in several iterations which we represented as rounds (Round-1, Round-2, Round-3, and Round-4). Each round served as a pilot study to determine the complexity of analysing the content of the images and generating high-quality annotation data. In each round, we executed different tasks (Task-1, Task-2, Task-3). These tasks represent only a fraction of potential concepts one can identify in the collection.

Task-1: involves the use of concrete food-related concepts. For the experiment, we selected a concrete concept “Fruit” and analysed the images by considering the presence/absence of fruit in the image. Fruit is selected due to its wider presence in the image collection and the concrete nature that makes it easier to be identified both by humans and existing CV tools with higher accuracy.

Task-2: focuses on images that contain abstract and subtle concepts that represent rich cultural aspects. In this task, we selected “Formal” and “Informal” concept categories and analysed the images by considering the setting where the food is presented.

Task-3: also focuses on the abstract and subjective concept that deals with appealing and non-appealing image categories. Definitions for images categorized as “Appealing” or “Non-appealing” depend on the overall aim of the project, after careful consultation of possible word definitions from online monolingual English dictionaries, such as Collins English Dictionary online, the American Heritage Dictionary and a book called “The Art Instinct”. In this respect, in our project we define an image that is “Appealing” as “an image that is a pleasure to look at”; an image that is “Non-appealing” is “an image that is not a pleasure to look at”. We are aware that these are highly subjective and will vary depending on other parameters such as cultural background, food preferences, etc.

The abstract categories, unlike the concrete concepts, embed some level of abstraction that can not be formally detected by both humans and computers due to a high level of subjectivity based on culture, dietary style, geographic location, and other states of the annotators.

This case study particularly focused on Appealing and Non-appealing images. We believe that the use of the Appealing/Non-appealing concept represents the majority of the desired semantic enrichment of cultural images due to the following reasons. First, ap-

pealingness can be defined based on the features of the images including colour, brightness, orientation, etc. [59]. Second, appealingness is subjective and it varies from one society to another society, in the time horizon and based on the dietary preference of individuals. This makes the concept very representative of the topics of cultural images. Since modern AI technologies are applied for cultural images, we wanted to explore how a CV model would understand an image that represents abstract concepts.

#### 4.2.2. Manual Annotation for Generating Training Data

During the preparation of the training data, selecting the initial sets of images for the semantic annotation process and generating high-quality data for training a CV system was crucial. The initial candidate images and the subsequent images used in the manual annotation are selected by the Europeana-local Austria experts by evaluating the appropriateness of the images for the task. During each round of tasks, new images that were not used in the previous rounds were added.

Five annotators were involved throughout the process and some additional annotators are introduced for further validation of the annotation results. The five annotators came from different educational backgrounds (digital humanist, semantic web expert, computer scientist, CV expert, student, and socio-linguist), geographic locations (Europe, South America, Africa, and Asia), gender (two female and three male) and dietary preference (vegans and vegetarians included). Although the diversity of the annotators is a certain factor, we did not make any scientific selection of these annotators to base any further analysis on the effect of their background on the annotation results. Table 2 presents the Kappa agreement between five annotators in Round-1 and Round-2 annotations. The results in the left column represent Kappa agreements from Round-1 where the annotators completed the tasks without consulting a formal definition of the annotation labels. The results in the right column present the Kappa agreements after the annotators are provided with a formal definition of all the categories. The effect of the presence of the formal definition can be compared in detail between Round-1 and Round-2 Kappa agreements for all three tasks.

**Table 2.** Inter-annotator agreement for Round-1 and Round-2 annotation.

Round-1					Round-2						
A001	A002	A003	A004	A005	A001	A002	A004	A005	A006		
Task-1: Fruit/Non-fruit.											
A001	1.000	0.928	0.892	0.907	0.886	A001	1.000	0.943	0.928	0.913	0.912
A002	0.928	1.000	0.892	0.938	0.897	A002	0.943	1.000	0.912	0.866	0.865
A003	0.892	0.892	1.000	0.923	0.923	A004	0.928	0.912	1.000	0.913	0.881
A004	0.907	0.938	0.923	1.000	0.918	A005	0.913	0.866	0.913	1.000	0.897
A005	0.886	0.897	0.923	0.918	1.000	A006	0.912	0.865	0.881	0.897	1.000
Task-2: Formal/Informal.											
A001	1.000	0.330	0.252	0.316	-0.091	A001	1.000	0.168	0.255	0.167	0.125
A002	0.330	1.000	0.210	0.306	0.153	A002	0.168	1.000	0.095	0.419	0.489
A003	0.252	0.210	1.000	0.051	-0.031	A004	0.255	0.095	1.000	0.089	0.208
A004	0.316	0.306	0.051	1.000	-0.028	A005	0.167	0.419	0.089	1.000	0.520
A005	-0.091	0.153	-0.031	-0.028	1.000	A006	0.125	0.489	0.208	0.520	1.000
Task-3: Appealing/Non-appealing.											
A001	1.000	0.659	0.296	0.534	0.317	A001	1.000	0.472	0.526	0.336	0.569
A002	0.659	1.000	0.325	0.453	0.268	A002	0.472	1.000	0.565	0.454	0.366
A003	0.296	0.325	1.000	0.424	0.370	A004	0.526	0.565	1.000	0.439	0.386
A004	0.534	0.453	0.424	1.000	0.454	A005	0.336	0.454	0.439	1.000	0.205
A005	0.317	0.268	0.370	0.454	1.000	A006	0.569	0.366	0.386	0.205	1.000

#### 4.2.3. Round-1

The first round image selection resulted in identifying 392 cultural images. The images were annotated by five annotators using Fruit/Non-fruit, Formal/Informal, and Appealing/Non-appealing categories. Annotators were asked to annotate the images without consulting any formal definitions of the categories. The resulting annotations were analysed and an inter-annotator agreement was generated. A higher level of kappa agreement (0.928) was achieved for Fruit/Non-fruit category and a fair agreement (0.317) was achieved in the Formal/Informal category, whereas moderate (0.528) agreement was achieved in the appealing/non-appealing category (Refer Table 2).

#### 4.2.4. Round-2

In the second round, we formally defined all the concepts to investigate the effect of having common semantics on the inter-annotator agreement. The definitions (refer Table 3) were provided to the annotators before they started Round-2 annotation.

**Table 3.** The definitions used for Round-2 ChIA image classification.

Concept	Definition
Fruit/Non-fruit	Fruit: a fruit is something that grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat. (e.g., strawberry, nut, tomato, peach, banana, green beans, melon, apple). Non-fruit: images that do not feature any type of fruit (for fruit definition see above)
Formal/Informal	Formal: arranged in a very controlled way or according to certain rules; an official situation or context. Informal: a relaxed environment, an unofficial situation or context, disorderly arrangement.
Appealing/Non-appealing	Appealing: an image that is a pleasure to look at. A food image that is pleasing to the eye, desirable to eat and good for food. Non-appealing: an image that is not a pleasure to look at.

The effect of the formal definition for concrete concepts (Fruit/Non-fruit) demonstrated little impact ( $0.928 \rightarrow 0.922$ ) as the concepts were clear and straightforward. However, the use of formal description of the abstract concepts showed a slight improvement in the inter-annotator agreement by providing more clarity about the features the annotators should consider during the annotation and enabled them to be more self-consistent Formal/Informal ( $0.31 \rightarrow 0.37$ ) and Appealing/Non-appealing ( $0.52 \rightarrow 0.50$ ) (note the decrease for Appealing/Non-appealing category). The results of the inter-annotator agreements are presented in Table 2. The definition for the Appealing/Non-appealing category was adopted in Round-3 and Round-4.

#### 4.2.5. Round-3

Based on the lesson learned in Round-2, we focused only on the Appealing/Non-appealing category. We dropped Fruit/Non-fruit category because existing CV tools have already achieved a higher level of accuracy in identifying fruits in digital images [51,60,61]. We further dropped the Formal/Informal category as it resulted in a lower agreement. We took the Appealing/Non-appealing category as it represents an interesting aspect of cultural images. In this round, all the annotations from all three rounds were used to train a CV model.

From our image collection, we selected 1010 additional images and included them in the annotation process. Of all the six annotators who participated in this round, five annotators were familiar with the process and the sixth annotator was included upon getting familiar with the process. The inter-annotator agreement between the annotators is presented in Table 4. Apart from the kappa agreement, we generated the number of images classified as appealing and non-appealing using majority voting methods. In this round,

we further build a CNN classifier using the collected annotation data as a data set. We used 830 images with greater than or equal to 66.7% vote. The annotation data was split into training, validation, and test sets. We trained three CNN models and identified the best performing model in Table 5. The Kappa agreement presented in the table shows that there is a fair agreement between the annotators on Round-3 images. The resulting models showed some promising results, however, the use of 830 images for training a model is not sufficient to make a reasonable conclusion. Another concern of using this data set as a basis for training a model was the difference between the number of Appealing and Non-appealing images, such imbalance created a bias in the CNN model, and hence to address this problem we selected additional 1079 images to run through Round-4 annotation.

**Table 4.** Inter-annotator agreement for Round-3 annotation of 1010 images (Appealing/Non-Appealing).

	A001	A002	A004	A005	A006	A007
A001	1.000	0.293	0.335	0.330	0.164	0.274
A002	0.293	1.000	0.475	0.483	0.190	0.042
A004	0.335	0.475	1.000	0.648	0.156	0.082
A005	0.330	0.483	0.648	1.000	0.123	0.061
A006	0.164	0.190	0.156	0.123	1.000	-0.025
A007	0.274	0.042	0.082	0.061	-0.025	1.000

**Table 5.** Results of training deep learning models for image classification: Round-3

Model	Training Accuracy	Validation Accuracy	Test Accuracy
Fine tuned ResNet50	83.51%	83.81%	80%
Fine tuned Inception_V3	92.61%	87.62%	90%
Fine tuned Xception	93.2%	88.1%	85.56%

#### 4.2.6. Round-4

Round-4 annotation aimed to increase the number of images for training and to balance the training data for the two categories (Appealing and Non-appealing). To achieve a high quality of the data set compared to the previous annotation round, the threshold was set to 80%. To achieve this, another 1079 images were added and the inter-annotator agreement is provided in Table 6. To build a balanced data set we reduced the number of images belonging to the Appealing category.

**Table 6.** Inter-annotator agreement for Round-4 annotation of 1079 images (Appealing/Non-Appealing).

	A001	A002	A004	A005	A006
A001	1.000	0.223	0.287	0.057	0.293
A002	0.223	1.000	0.245	0.090	0.163
A004	0.287	0.245	1.000	0.133	0.200
A005	0.057	0.090	0.133	1.000	0.085
A006	0.293	0.163	0.200	0.085	1.000

Once again, deep learning models were trained using 1010 images of which 545 images belonged to the Appealing category and 465 images belonged to the Non-appealing category. This image data was further divided into training, validation, and test set. There was an improvement in the distribution of images belonging to both categories in the final data. In combination with previously trained models on a huge data set, our image collection provides a very good set of training data with a reasonable annotation accuracy. The obtained data shows that compared to the human annotation, the accuracy of the CV models is superior in that, first, it always generates a consistent label (we observed that

human annotators were not always consistent between the rounds), and second, it can be trained using active learning where the models learn by incorporating feedback from users.

In this round, we decided to include a confidence score of the CV models regarding the predicted category of a target image. This is a significant step that enables us to incorporate subjectivity as views related to a given culture are subjective. Detailed results of the CV models implemented are provided in Table 7. The best performing model in our case is the Fine-tuned Xception model which outperformed the other two models.

**Table 7.** Results of training deep learning models for image classification: Round-4

Model	Training Accuracy	Validation Accuracy	Test Accuracy
Fine tuned ResNet50	87.47%	84.5%	83.85%
Fine tuned Inception_V3	81.68%	85.5%	88.46%
Fine tuned Xception	95.98%	88.5%	90.77%

Although the four rounds enabled us to raise the quality and quantity of the training data sets, CV models benefit from large training data sets. As we demonstrated in our experiment, generating a large and high-quality data set in the cultural domain requires a huge resource, particularly the availability of expert annotators. However, our approach provides a methodology that ensures the incremental generation of high-quality training data set for domains with low resources.

#### 4.3. Phase-4: Integration and Exploitation Phase

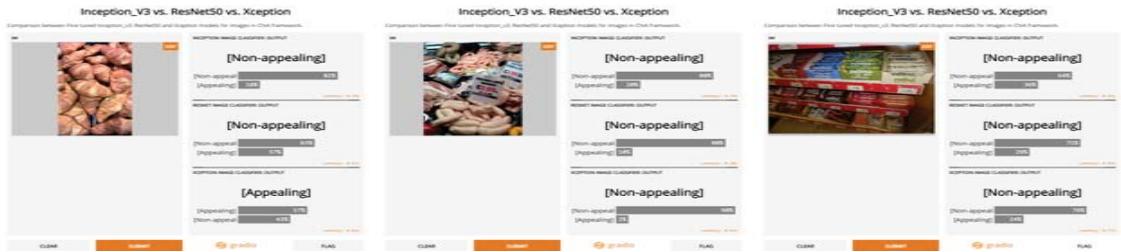
An important aspect of semantic annotation is the integration of the resulting annotation into an existing repository in a form that is suitable for both humans and machines to understand and interpret. The integration process first converts the resulting semantic annotation into s-p-o triples as discussed in Section 3.3.1. Second, integrating the newly generated triples into the existing semantic repository, and third, supporting efficient exploration of the data and exploitation of the images by the end-users. Each step is discussed as follows.

##### 4.3.1. Moving towards Large Scale Annotation

The next step is the application of the trained models to predict the labels for the new and unseen images. At this stage, we have identified and trained three models with 81.68% to 95.98% training accuracy, and 84.5% to 88.5% validation accuracy. An image is classified and annotated with a particular class label by considering the average of the confidence level of the majority voted class, a similar approach has been followed in [62]. The prediction of some selected images using the three models along with the confidence scores is presented in Figure 3.



**Figure 3. Cont.**



**Figure 3.** Sample images and their predicted categories using the three models. The images are cultural food images that are taken from the Europeana platform. The predictions of the three models indicate the categories (Appealing/Non-appealing) of the images along with the confidence score of each model.

#### 4.3.2. Integration of Results

The annotations and their respective confidence scores are used to create s-p-o triples. These triples are generated for every image and the resulting data can be integrated into existing platforms following the preference of the aggregators. These data sets can be pushed to any triple store (including Europeana-local Austria) once verified by the aggregators. In this semantic interlinking stage, we link the images with concepts drawn from an ontology related to emotion (<http://www.ontobee.org/ontology/MFOEM> (accessed on 15 May 2021)) using a rdfs:type property. The images are also labelled as rdfs:type edm:WebResource. We further add the labels as part of the metadata using dc:description and rdfs:comment to represent it as a free-text account of the image resources. It is also observed that without a specialised ontology, the accurate interlinking of the annotation data with existing ontologies will not fully meet the objectives. Aggregators can link the target images using relationships and concepts by including additional triples using the R2RML mapping discussed in Section 3.3.1. For brevity, we present a snippet of the generated triples below.

```

<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24255> a
<http://www.europeana.eu/schemas/edm/webResource> ;
<http://www.w3.org/2000/01/rdf-schema#comment> "Appealing" , "Appealing:95.69" ;
<http://purl.org/dc/elements/1.1/description> "Appealing" , "Appealing:95.69" ;
<http://purl.org/dc/elements/1.1/subject> <http://purl.obolibrary.org/obo/MFOEM_000039> .
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24264> a
<http://www.europeana.eu/schemas/edm/webResource> ;
<http://www.w3.org/2000/01/rdf-schema#comment> "Appealing" , "Appealing:96.33" ;
<http://purl.org/dc/elements/1.1/description> "Appealing" , "Appealing:96.33" ;
<http://purl.org/dc/elements/1.1/subject> <http://purl.obolibrary.org/obo/MFOEM_000039> .
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24245> a
<http://www.europeana.eu/schemas/edm/webResource> ;
<http://www.w3.org/2000/01/rdf-schema#comment> "Non-appealing" , "Non-appealing:81.24" ;
<http://purl.org/dc/elements/1.1/description> "Non-appealing" , "Non-appealing:81.24" .
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24263> a
<http://www.europeana.eu/schemas/edm/webResource> ;
<http://www.w3.org/2000/01/rdf-schema#comment> "Appealing" , "Appealing:91.9" ;
<http://purl.org/dc/elements/1.1/description> "Appealing" , "Appealing:91.9" ;
<http://purl.org/dc/elements/1.1/subject> <http://purl.obolibrary.org/obo/MFOEM_000039> .

```

#### 4.3.3. Supporting Efficient Exploration

The methodology further provided mechanisms for efficient exploration of the resources by enabling exploration paths and templates. The following SPARQL templates are introduced to support the explorations [47]. The exploration of the triples is not restricted to the exploration paths, however becomes open and can be used for interlinking images with selected abstract cultural queries. Sample SPARQL query for extracting Appealing (Aesthetically Pleasing) images by filtering the subject using rdf:type and dc:subject predicates is shown below.

```

prefix obo: <http://purl.obolibrary.org/obo/>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix dc: <http://purl.org/dc/elements/1.1/>
prefix edm: <http://www.europeana.eu/schemas/edm/>

select ?subject ?predicate ?object
where{
?subject ?predicate ?object.
?subject rdf:type edm:webResource.
?subject dc:subject obo:MFOEM_000039.
}
limit 15

```

A snippet of the output of the above query is given below. The result shows the potential of the newly generated metadata to enrich, interlink and group images by tagging them using one or more abstract cultural heritage concepts.

```

<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24255> dc:subject obo:MFOEM_000039
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24255> dc:description Appealing
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24255> dc:description Appealing:95.69
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24255> rdfs:comment Appealing
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24255> rdfs:comment Appealing:95.69
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24255> rdf:type edm:webResource
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24264> dc:subject obo:MFOEM_000039
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24264> dc:description Appealing
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24264> dc:description Appealing:96.33
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24264> rdfs:comment Appealing
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24264> rdfs:comment Appealing:96.33
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24264> rdf:type edm:webResource
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24263> dc:subject obo:MFOEM_000039
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24263> dc:description Appealing
<https://www.europeana.eu/en/item/2059511/data_foodanddrink_24263> dc:description Appealing:91.9

```

The application of our method to represent and annotate cultural images using abstract concepts is scalable when additional cultural concepts are used to annotate the target images. Depending on the requirements, it supports the extraction of images linked to external repositories.

## 5. Discussion

Even if the individual phases and steps proposed in our methodology are not new, this paper presents a novel and efficient combination of the steps that fit the purpose. The application of the methodology in the case study on Europeana cultural heritage images exposed the problems and the challenges not only from the methodological but also technical and practical perspectives. This paper implements the methodology, first by applying ontologies to consistently represent concepts, second, using CV tools to enrich cultural images by training models that can be applied to large data sets, third, by enabling existing systems to efficiently support user requirements, and finally integrating subjectivity and fuzziness into the metadata. Lessons learned from each of the contributions are summarised below.

Firstly, although several ontologies and vocabularies are available, the selection and composition of ontologies that represent the knowledge base of cultural and historical aspects are not fully explored. The composition of cultural heritage concepts from existing generic or specific ontologies is difficult and deserves a proper investigation. It is evident from our case study that several cultural aspects (including family status, economic status, style, nutrition, etc.) are embedded in the images which need to be formally defined using vocabularies. As an example, we searched Linked Open Vocabulary (LOV (<https://lov.linkeddata.es/dataset/lov/> (accessed on 15 July 2021))) repository for concepts representing appealingness and attractiveness defined concerning food images. Our search resulted in a few concepts that are only related to computer software quality features. All these aspects can not be covered in a single ontology, therefore the integration of several existing ontologies and the development of new ones is crucial. In addition to the vocabularies, domain-specific relationships between the images and the concepts need to be defined. We came across cases where the annotated features of the images can not be

embedded using existing generic relations. In this regard, ontologies play a major role in defining rich relationships (`owl:ObjectProperty`) between cultural concepts.

Secondly, CV tools have provided significant breakthroughs in detecting objects. Although they lag in identifying exceptional cases. Our case study demonstrates that they can be effectively exploited. The case study expanded the state of the art by including the detection of abstract concepts that are very subjective and difficult to quantify. We are aware that there are significant omissions of exceptional cases by the AI and ML algorithms [63] and tried to reduce the bias by incorporating confidence scores. We are also aware that the kappa inter-annotator agreement may cause an issue in the interpretation of the agreements as poor, slight, fair, moderate, substantial, and strong [64,65]. One limitation of our method is that the use of Kappa agreement and its interpretation which may not be suitable for mission-critical tasks such as in medical applications [66,67]. To reduce the effect of outlier cases, we embedded the confidence levels of the predicted annotations. After all, for such abstract concepts, the experiment also showed that the inter-annotator agreements are low or moderate compared to that of concrete concepts. Round-2 experiment enabled us to identify some interesting aspects of cultural image enrichment. Concrete concepts can be annotated by existing image recognition tools with high accuracy [68], whereas abstract concepts are fuzzy even when the annotation is done by human experts.

Thirdly, the digital humanities domain, particularly cultural heritage could benefit from existing semantic web and AI domains in several ways. However, our experiment also showed that there is a lot of work that needs to be done to ensure the quality during the digitization process of cultural heritage resources. The integration of new annotations on top of existing annotations should not introduce an inconsistent interpretation of the target resources.

Finally, our research contributes additional data sets to the research community. The data set includes more than 2000 images that are annotated by five annotators. This data set can be used as a benchmark for evaluating future models and also serve as a starting point for future crowd annotation. Our food vocabulary is another contribution to the domain in that it amalgamates food concepts from different sources into one.

## 6. Conclusions

We presented a methodology for semantic enrichment of digital cultural heritage images covering the domain of cultural food images. A three-phase methodology is proposed and a case study following the methodology is implemented in the context of a 2-year ChIA project. The proposed methodology provides a structured approach that enables digital humanities experts to identify, enrich and publish their cultural heritage-related collections using LOD formats. It also provides guidance and directions on how existing artificial intelligence tools such as CV and semantic web technologies can be combined and exploited efficiently in the digital humanities domain. This paper explored the introduction of abstract and subjective concepts into the proposed semantic enrichment process and identified the challenges and opportunities that exist in the emerging AI-based technologies. We believe that the confidence of the CV model can be continually improved by incorporating active learning steps in the methodology. In future work, we will integrate an active learning component, as discussed in [15], where users will be provided options to rate the output while the model continually improves its performance by learning from the feedback. Another future research will be contributing towards an ontology of abstract cultural concepts and the preparation of high-quality data sets that can be used in similar research settings.

**Author Contributions:** Conceptualization, Y.A., J.M.; methodology, Y.A., R.R.S.; software, Y.A., R.R.S. and J.M.; validation, Y.A., R.R.S. and J.M.; formal analysis, R.R.S., J.M. and Y.A.; investigation, Y.A., R.R.S., J.M., G.K. and A.D.; resources, Y.A., R.R.S., J.M., G.K. and A.D.; data curation, G.K., Y.A.; writing—original draft preparation, Y.A.; writing—review and editing, J.M., G.K., A.D. and Y.A.; visualization, Y.A., J.M.; supervision, Y.A.; project administration, A.D., Y.A.; funding acquisition, Y.A., A.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Austrian Academy of Sciences [Österreichische Akademie der Wissenschaften] go!digital Next Generation grant (GDNG 2018-051).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The image data used in this research is available at the ChIA website (<https://chia.acdh.oeaw.ac.at/data/>) (accessed on 15 July 2021). The data set includes the URLs to the images along with annotations from all rounds. data set predicting the labels for additional images is also included. The ChIA vocabulary is available at Europeana local-At (<http://chia.ait.co.at/vocab/ChIA/index.php>) (accessed on 15 July 2021)). Finally, the three trained models are also available as IPython Notebooks (<https://github.com/acdh-oeaw/Chia/blob/master/notebooks/nnetworks/>) (accessed on 15 July 2021)).

**Acknowledgments:** This research is supported by ADAPT and the Austrian Academy of Sciences go!digital Next Generation grant (GDNG 2018-051). The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and isco-funded under the European Regional Development Fund(ERDF) through Grant # 13/RC/2106.

**Conflicts of Interest:** There is no conflict of interest.

## References

- Ziku, M. Digital Cultural Heritage and Linked Data: Semantically-informed conceptualisations and practices with a focus on intangible cultural heritage. *Liber Q.* **2020**, *30*. [[CrossRef](#)]
- Meroño-Peñuela, A.; Ashkpour, A.; Van Erp, M.; Mandemakers, K.; Breure, L.; Scharnhorst, A.; Schlobach, S.; Van Harmelen, F. Semantic Technologies for Historical Research: A Survey. *Semant. Web* **2015**, *6*, 539–564. [[CrossRef](#)]
- Beretta, F.; Ferhod, D.; Gedzelman, S.; Vernus, P. The SyMoGIH project: Publishing and sharing historical data on the semantic web. In Proceedings of the Digital Humanities 2014, Lausanne, Switzerland, 8–12 July 2014; pp. 469–470.
- Doerr, M. Ontologies for Cultural Heritage. In *Handbook on Ontologies*; International Handbooks on Information Systems; Staab, S., Studer, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 463–486. [[CrossRef](#)]
- Fiorucci, M.; Khoroshiltsseva, M.; Pontil, M.; Travaglia, A.; Del Bue, A.; James, S. Machine Learning for Cultural Heritage: A Survey. *Pattern Recognit. Lett.* **2020**, *133*, 102–108. [[CrossRef](#)]
- Evens, T.; Hauttekeete, L. Challenges of digital preservation for cultural heritage institutions. *J. Librariansh. Inf. Sci.* **2011**, *43*, 157–165. [[CrossRef](#)]
- Hyvönen, E. Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. *Semant. Web* **2020**, *11*, 187–193. [[CrossRef](#)]
- Cornia, M.; Stefanini, M.; Baraldi, L.; Corsini, M.; Cucchiara, R. Explaining digital humanities by aligning images and textual descriptions. *Pattern Recognit. Lett.* **2020**, *129*, 166–172. [[CrossRef](#)]
- Cosovic, M.; Jankovic, R.; Ramic-Brkic, B. Cultural Heritage Image Classification. In *Data Analytics for Cultural Heritage: Current Trends and Concepts*; Belhi, A., Bouras, A., Al-Ali, A.K., Sadka, A.H., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 25–45. [[CrossRef](#)]
- Janković, R. Machine Learning Models for Cultural Heritage Image Classification: Comparison Based on Attribute Selection. *Information* **2020**, *11*, 12. [[CrossRef](#)]
- Ciocca, G.; Napoletano, P.; Schettini, R. CNN-based features for retrieval and classification of food images. *Comput. Vis. Image Underst.* **2018**, *176–177*, 70–77. [[CrossRef](#)]
- Gruber, T. Ontology. In *Encyclopedia of Database Systems*; Liu, L., Özsü, M.T., Eds.; Springer: Boston, MA, USA, 2009; pp. 1963–1965. [[CrossRef](#)]
- Zeng, M.L. Semantic enrichment for enhancing LAM data and supporting digital humanities. Review article. *Prof. Inf.* **2019**, *28*. [[CrossRef](#)]
- Lei, X.; Meroño-Peñuela, A.; Zhisheng, H.; van Harmelen, F. An Ontology Model for Narrative Image Annotation in the Field of Cultural Heritage. In Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe), Vienna, Austria, 21–25 October 2017.
- Musik, C.; Zeppelzauer, M. Computer Vision and the Digital Humanities: Adapting Image Processing Algorithms and Ground Truth through Active Learning. *View J. Eur. Telev. Hist. Cult.* **2018**, *7*, 59–72. [[CrossRef](#)]
- Triandis, H. Subjective Culture. *Online Read. Psychol. Cult.* **2002**, *2*. [[CrossRef](#)]
- Zhu, X.; Vondrick, C.; Fowlkes, C.C.; Ramanan, D. Do we need more training data? *Int. J. Comput. Vis.* **2016**, *119*, 76–92. [[CrossRef](#)]
- Dorn, A.; Abgaz, Y.; Koch, G.; Díaz, J.L.P. Harvesting Knowledge from Cultural Images with Assorted Technologies: The Example of the ChIA Project. In *Knowledge Organization at the Interface: Proceedings of the Sixteenth International ISKO Conference*,

- 2020 *Aalborg, Denmark*, 1st ed.; International Society for Knowledge Organization (ISKO); Lykke, M., Svarre, T., Skov, M., Martínez-Ávila, D., Eds.; Ergon-Verlag: Baden, Germany, 2020; pp. 470–473. [[CrossRef](#)]
19. Sorbara, A. Digital Humanities and Semantic Web. The New Frontiers of Transdisciplinary Knowledge. In *Expanding Horizons: Business, Management and Technology for Better Society*; ToKnowPress: Bangkok, Thailand, 2020; p. 537.
20. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [[CrossRef](#)]
21. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
22. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [[CrossRef](#)]
24. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [[CrossRef](#)]
25. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [[CrossRef](#)]
26. Berners-Lee, T.; Hendler, J.; Lassila, O. The semantic web. *Sci. Am.* **2001**, *284*, 34–43. [[CrossRef](#)]
27. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*; Aberer, K., Choi, K.S., Noy, N., Allemang, D., Lee, K.I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 722–735.
28. Navigli, R.; Ponzetto, S.P. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.* **2012**, *193*, 217–250. [[CrossRef](#)]
29. Miller, G.A. WordNet: A Lexical Database for English. *Commun. ACM* **1995**, *38*, 39–41. [[CrossRef](#)]
30. Doerr, M. The CIDOC CRM—An Ontological Approach to Semantic Interoperability of Metadata. *AI Mag. AIM* **2003**, *24*, 75–92. [[CrossRef](#)]
31. Isaac, A. *Europeana Data Model Primer*; Technical Report; European Commission: Brussels, Belgium, 2013.
32. Meroño-Peñuela, A. Digital Humanities on the Semantic Web: Accessing Historical and Musical Linked Data. *J. Catalan Intellect. Hist.* **2017**, *1*, 144–149. [[CrossRef](#)]
33. Borgman, C. The Digital Future is Now: A Call to Action for the Humanities. *Digit. Humanit. Q.* **2010**, *3*, 1–30.
34. Commission, E. *Commission Recommendation of 27.10.2011 on the Digitisation and Online Accessibility of Cultural Material and Digital Preservation*; European Commission: Brussels, Belgium, 2017.
35. Sabou, M.; Lopez, V.; Motta, E.; Uren, V. Ontology selection: Ontology evaluation on the real Semantic Web. In Proceedings of the 15th International World Wide Web Conference (WWW 2006), Edinburgh, UK, 23–26 May 2006.
36. Pileggi, S.F. Probabilistic Semantics. *Procedia Comput. Sci.* **2016**, *80*, 1834–1845. [[CrossRef](#)]
37. Rocha Souza, R.; Dorn, A.; Piringer, B.; Wandl-Vogt, E. Towards A Taxonomy of Uncertainties: Analysing Sources of Spatio-Temporal Uncertainty on the Example of Non-Standard German Corpora. *Informatics* **2019**, *6*, 34. [[CrossRef](#)]
38. Chen, M.; Dai, W.; Sun, S.Y.; Jonasch, D.; He, C.Y.; Schmid, M.F.; Chiu, W.; Ludtke, S.J. Convolutional neural networks for automated annotation of cellular cryo-electron tomograms. *Nat. Methods* **2017**, *14*, 983–985. [[CrossRef](#)]
39. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [[CrossRef](#)]
40. Abgaz, Y.; Dorn, A.; Piringer, B.; Wandl-Vogt, E.; Way, A. Semantic Modelling and Publishing of Traditional Data Collection Questionnaires and Answers. *Information* **2018**, *9*, 297. [[CrossRef](#)]
41. Abgaz, Y.; Dorn, A.; Piringer, B.; Wandl-Vogt, E.; Way, A. A semantic model for traditional data collection questionnaires enabling cultural analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*; McCrae, J.P., Chiarcos, C., Declerck, T., Gracia, J., Klimek, B., Eds.; European Language Resources Association (ELRA): Paris, France, 2018.
42. Jones, D.; O'Connor, A.; Abgaz, Y.M.; Lewis, D. A Semantic Model for Integrated Content Management, Localisation and Language Technology Processing. In *Proceedings of the 2nd International Conference on Multilingual Semantic Web*; DEU: Aachen, Germany, 2011; Volume 775, pp. 38–49.
43. Dorn, A.; Wandl-Vogt, E.; Abgaz, Y.; Benito Santos, A.; Therón, R. Unlocking cultural conceptualisation in indigenous language resources: Collaborative computing methodologies. In *Proceedings of the LREC 2018 Workshop CCURL 2018*, Miyazaki, Japan, 7–12 May 2018.
44. Debruyne, C.; O'Sullivan, D. R2RML-F: Towards Sharing and Executing Domain Logic in R2RML Mappings. In *Proceedings of the Workshop on Linked Data on the Web, LDOW 2016, Co-Located with 25th International World Wide Web Conference (WWW 2016)*; CEUR Workshop Proceedings; Auer, S., Berners-Lee, T., Bizer, C., Heath, T., Eds.; RWTH: Aachen, Germany, 2016; Volume 1593.
45. Isaac, A.; Haslhofer, B. Europeana linked open data—data. europeana. eu. *Semant. Web* **2013**, *4*, 291–297. [[CrossRef](#)]

46. Schreiber, G.; Amin, A.; Arroyo, L.; van Assem, M.; de Boer, V.; Hardman, L.; Hildebrand, M.; Omelayenko, B.; van Osenbruggen, J.; Tordai, A.; et al. Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator. *J. Web Semant.* **2008**, *6*, 243–249. [[CrossRef](#)]
47. Rodríguez Díaz, A.; Benito-Santos, A.; Dorn, A.; Abgaz, Y.; Wandl-Vogt, E.; Therón, R. Intuitive Ontology-Based SPARQL Queries for RDF Data Exploration. *IEEE Access* **2019**, *7*, 156272–156286. [[CrossRef](#)]
48. Ait-Mlouk, A.; Jiang, L. KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding Over Linked Data. *IEEE Access* **2020**, *8*, 149220–149230. [[CrossRef](#)]
49. Al-Zubaide, H.; Issa, A.A. OntBot: Ontology based chatbot. In Proceedings of the International Symposium on Innovations in Information and Communications Technology, Amman, Jordan, 29 November–1 December 2011; pp. 7–12. [[CrossRef](#)]
50. Abgaz, Y.; Dorn, A.; Preza Diaz, J.L.; Koch, G. Towards a Comprehensive Assessment of the Quality and Richness of the Europeana Metadata of food-related Images. In Proceedings of the 1st International Workshop on Artificial Intelligence for Historical Image Enrichment and Access, Marseille, France, 11–16 May 2020; European Language Resources Association (ELRA): Marseille, France, 2020; pp. 29–33.
51. Preza Diaz, J.L.; Dorn, A.; Koch, G.; Abgaz, Y. A comparative approach between different Computer Vision tools, including commercial and open-source, for improving cultural image access and analysis. In Proceedings of the The 10th International Conference on Advanced Computer Information Technologies (ACIT’2020), Deggendorf, Germany, 16–18 September 2020; [[CrossRef](#)]
52. Leatherdale, D.; Tidbury, G.E.; Mack, R.; Food and Agriculture Organization of the United Nations; Commission of the European Communities. *AGROVOC: A Multilingual Thesaurus of Agricultural Terminology*, english version ed.; Apimondia, by arrangement with the Commission of the European Communities: Rome, Italy, 1982; p. 530.
53. Alexiev, V. Museum linked open data: Ontologies, datasets, projects. *Digit. Present. Preserv. Cult. Sci.* **2018**, *VIII*, 19–50.
54. Petersen, T. Developing a new thesaurus for art and architecture. *Libr. Trends* **1990**, *38*, 644–658.
55. Molholt, P.; Petersen, T. The role of the ‘Art and Architecture Thesaurus’ in communicating about visual art. *Ko Knowl. Organ.* **1993**, *20*, 30–34. [[CrossRef](#)]
56. Baca, M.; Gill, M. Encoding multilingual knowledge systems in the digital age: The getty vocabularies. *NASKO* **2015**, *5*, 41–63. [[CrossRef](#)]
57. Alghamdi, D.A.; Dooley, D.M.; Gosal, G.; Griffiths, E.J.; Brinkman, F.S.; Hsiao, W.W. *FoodOn: A Semantic Ontology Approach for Mapping Foodborne Disease Metadata*; ICBO: Lansing, MI, USA, 2017.
58. Popovski, G.; Korousic-Seljak, B.; Eftimov, T. FoodOntoMap: Linking Food Concepts across Different Food Ontologies. In Proceedings of the KEOD, Vienna, Austria, 17–19 September 2019; pp. 195–202.
59. Toet, A.; Kaneko, D.; de Kruijf, I.; Ushijima, S.; van Schaik, M.G.; Brouwer, A.M.; Kallen, V.; van Erp, J.B.F. CROCUFID: A Cross-Cultural Food Image Database for Research on Food Elicited Affective Responses. *Front. Psychol.* **2019**, *10*, 58. [[CrossRef](#)] [[PubMed](#)]
60. Zawbaa, H.M.; Abbass, M.; Hazman, M.; Hassenian, A.E. Automatic Fruit Image Recognition System Based on Shape and Color Features. In *Advanced Machine Learning Technologies and Applications*; Hassanien, A.E., Tolba, M.F., Taher Azar, A., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 278–290.
61. Bresilla, K.; Perulli, G.D.; Boini, A.; Morandi, B.; Corelli Grappadelli, L.; Manfrini, L. Single-Shot Convolution Neural Networks for Real-Time Fruit Detection Within the Tree. *Front. Plant Sci.* **2019**, *10*, 611. [[CrossRef](#)]
62. Xue, D.; Zhou, X.; Li, C.; Yao, Y.; Rahaman, M.M.; Zhang, J.; Chen, H.; Zhang, J.; Qi, S.; Sun, H. An Application of Transfer Learning and Ensemble Learning Techniques for Cervical Histopathology Image Classification. *IEEE Access* **2020**, *8*, 104603–104618. [[CrossRef](#)]
63. Birhane, A. The Impossibility of Automating Ambiguity. *Artif. Life* **2021**, *1*–18. [[CrossRef](#)]
64. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [[CrossRef](#)]
65. Cohen, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213–220. [[CrossRef](#)]
66. McHugh, M. Interrater reliability: The kappa statistic. *Biochem. Med. Cas. Hrvat. Drus. Med. HDMB* **2012**, *22*, 276–282. [[CrossRef](#)]
67. Tang, W.; Hu, J.; Zhang, H.; Wu, P.; He, H. Kappa coefficient: A popular measure of rater agreement. *Shanghai Arch. Psychiatry* **2015**, *27*, 62–67. [[CrossRef](#)] [[PubMed](#)]
68. Ciocca, G.; Napoletano, P.; Schettini, R. Food Recognition: A New Dataset, Experiments, and Results. *IEEE J. Biomed. Health Inform.* **2017**, *21*, 588–598. [[CrossRef](#)] [[PubMed](#)]



Article

# ChainLineNet: Deep-Learning-Based Segmentation and Parameterization of Chain Lines in Historical Prints

Aline Sindel<sup>1,\*</sup>, Thomas Klinke<sup>2</sup>, Andreas Maier<sup>1</sup> and Vincent Christlein<sup>1</sup>

<sup>1</sup> Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91058 Erlangen, Germany; andreas.maier@fau.de (A.M.); vincent.christlein@fau.de (V.C.)

<sup>2</sup> Cologne Institute of Conservation Sciences (CICS), Technische Hochschule Köln, 50678 Köln, Germany; thomas.klinke@th-koeln.de

\* Correspondence: aline.sindel@fau.de

**Abstract:** The paper structure of historical prints is sort of a unique fingerprint. Paper with the same origin shows similar chain line distances. As the manual measurement of chain line distances is time consuming, the automatic detection of chain lines is beneficial. We propose an end-to-end trainable deep learning method for segmentation and parameterization of chain lines in transmitted light images of German prints from the 16th Century. We trained a conditional generative adversarial network with a multitask loss for line segmentation and line parameterization. We formulated a fully differentiable pipeline for line coordinates' estimation that consists of line segmentation, horizontal line alignment, and 2D Fourier filtering of line segments, line region proposals, and differentiable line fitting. We created a dataset of high-resolution transmitted light images of historical prints with manual line coordinate annotations. Our method shows superior qualitative and quantitative chain line detection results with high accuracy and reliability on our historical dataset in comparison to competing methods. Further, we demonstrated that our method achieves a low error of less than 0.7 mm in comparison to manually measured chain line distances.



**Citation:** Sindel, A.; Klinke, T.; Maier, A.; Christlein, V. ChainLineNet: Deep-Learning-Based Segmentation and Parameterization of Chain Lines in Historical Prints. *J. Imaging* **2021**, *7*, 120. <https://doi.org/10.3390/jimaging7070120>

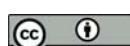
Academic Editors: Giovanna Castellano, Gennaro Vessio and Fabio Bellavia

Received: 4 June 2021

Accepted: 13 July 2021

Published: 19 July 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



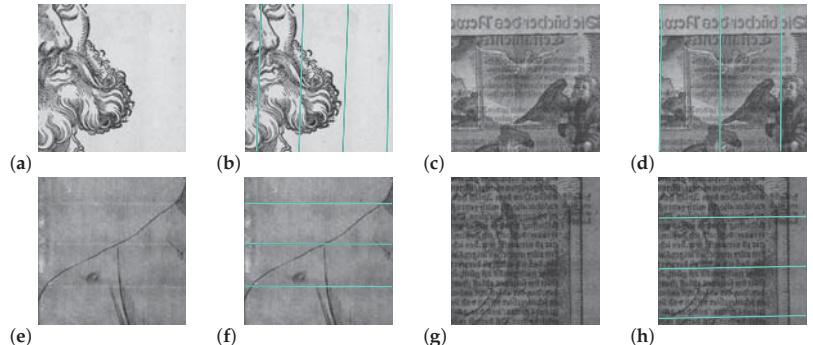
**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Since ancient times, paper has played a prominent role as a carrier for information. In the 16th Century, the only available paper was laid paper, which was manually produced in paper mills. Wood, old rags, and other ingredients were stamped and macerated in water into a pulp of fibers. Then, the paper was scooped by hand using a mold with a wire sieve made of closely spaced "laid" wires and perpendicular more widely spaced "chain" wires. After scooping the fibers from the vat, the remaining fibrous web on the wire sieve forms the paper [1]. On its surface, the grid pattern of the wires is imparted, as can be seen in the transmitted light photographs in Figure 1a,c,e,g. In addition, a watermark can be embedded into the paper structure as a seal of quality and origin by placing bent metal wires on the sieve. Concerning laid paper, the distances between the parallel chain lines vary across the sieve, but are approximately 25–30 mm [2]. For every mold, the chain lines form a unique pattern. Papers created by the same mold show a similar pattern of chain line distances. The impression of the sieve provides a unique conclusion to identify the mold. Images formed by the same mold are called moldmates [1]. Papers from different origins have different line sequences. Characteristics of the paper structure, such as the shape and placement of watermarks, chain line intervals, and the density of laid lines provide possibilities for computer vision to support art historical research. Apart from analyzing the motif itself, e.g., concerning the degree of wear, also, chain line distances can give hints about dating, author assignment, and the chronology of writings and prints [3]. For further refinements, chain line intervals can be analyzed in combination with the density of laid

lines, watermarks, and histological findings on the fibers. Traditionally, chain line distances are manually measured by art technologists during the examination and visual inspection of the individual prints, which is very time consuming.

In this paper, we propose an end-to-end trainable method for segmentation and parameterization of chain lines in transmitted light images of German prints from the 16th Century. Our method exploits the power of deep neural networks in combination with prior knowledge from image and signal processing. We trained a conditional generative adversarial network by using a multitask loss for line segmentation and line parameterization. For the estimation of line coordinates, we designed a fully differentiable pipeline that comprises the steps of line segmentation, horizontal alignment and 2D Fourier filtering of line segments, line region proposals, and differentiable line fitting. For training and evaluation, we created a dataset of high-resolution transmitted light images of historical prints for which we manually annotated line coordinates. Our ChainLineNet learns to detect the chain lines with high reliability even if there are interferences caused by watermarks or if the lines are partly occluded by the ink of the artwork; cf. Figure 1.



**Figure 1.** The paper structure in historical prints consists of chain and laid lines, which are perpendicular to each other. Examples using transmitted light photography (a,c) for vertical and (e,g) for horizontal chain lines are shown. Our ChainLineNet effectively detects the chain lines (b,d,f,h); even so, these are partly occluded by the ink of the artworks. Detail images: (a) Hans Sebald Beham, *Martin Luther as Junker Jörg*, Woodcut, Germanisches Nationalmuseum Nürnberg, H1933; (c) Unknown, *Martin Luther*, Woodcut, Landesbibliothek Coburg, P I 6/12; (e) Lucas Cranach the Elder, *Martin Luther as Junker Jörg*, Woodcut, Klassik Stiftung Weimar, Bestand Museen, DK 181/83; (g) Hans Baldung Grien, *Martin Luther as Augustinian monk*, Woodcut, Klassik Stiftung Weimar, Herzogin Anna Amalia Bibliothek, Aut. Luther 1520:64; images captured by Thomas Klinke; all rights reserved by the respective museum/library.

## 2. Related Work

To digitize the paper structure of historical prints, several imaging techniques, e.g., beta-radiography, transmitted light photography, transmitted infrared, or thermography, can be applied. Transmitted light photography is a very fast application, inexpensive, and very easy to handle. Hence, additional image processing might be necessary due to interferences such as ink that remain visible. These interferences disappear in the images using the other modalities, but especially beta-radiography is only applicable for large institutions due to the necessary technical and financial input.

### 2.1. Segmentation and Detection of Chain Lines

There are a few approaches for the automated segmentation of chain lines of paper. Van der Lubbe et al. [3] assumed straight and vertical chain lines for chain line detection in radiography. They used uniform filtering and morphological opening and closing operators as the preprocessing and applied a vertical projection to detect the vertical lines as peaks of the projection. Atanasiu [4] proposed a software measurement tool to analyze the density

of laid lines by using the bidimensional discrete fast Fourier transform. In a preprocessing step, an emboss edge-enhancing high-pass filter reduces noise; however, the orientation of the laid lines has to be determined beforehand. Van Staalduin et al. [5] presented an approach for moldmate matching using the specific paper features of chain and laid lines. The lines are detected by means of the shadow around the chain lines. The sequences of line distances for moldmate matching are computed with a combination of the discrete Fourier transform and Radon transform based on the assumption of straight and equidistant lines. Hiary et al. [2] focused on the digitization, extraction, and graphical representation of watermarks. They used backlight scanning and image processing such as mathematical morphological operations to automatically extract and convert watermarks to graphical representations. In an intermediate step, they rotated the image to upright the chain lines by means of chain line detection and Radon transform. Johnson et al. [1] published a method to find moldmates among Rembrandt's prints in beta-radiographs. Their chain line pattern matching approach uses unique chain spacing sequences in the paper structure rather than watermarks to identify the moldmates. Based on the assumption of straight, but not necessarily parallel lines, they rotated the chain lines to the vertical and obtained the angle of rotation by applying the Radon transform. Finally, the lines were detected using a vertical filter and the Hough transform.

In our previous work [6], we trained a convolutional neural network (CNN) to automatically segment the chain lines in artworks. Therefore, we employed the UNet [7] as the network architecture and proposed two postprocessing steps by employing either random sample consensus (RANSAC) [8] or the Hough transform to locate and parameterize complete lines in the binarized segmentation results. First, we determined the global orientation of the lines (horizontal or vertical) based on applying the Sobel filter. For the RANSAC-based approach, we extracted line segments from the segmentation mask using connected components and filtered out too small or falsely oriented line segments. The remaining line segments were grouped using agglomerative clustering, and RANSAC was utilized to fit lines through each group of points. For the Hough-based approach, we applied Hough voting on the segmentation masks and used agglomerative clustering to merge line predictions.

## 2.2. Segmentation and Detection of Lines

Looking more generally at the task of line detection in the fields of wireframe detection and semantic and horizon line detection, deep learning has been extensively applied.

Wireframe detection is the detection of line segments and junctions in a scene to describe all kinds of geometric objects or architectures [9]. Huang et al. [9] proposed a two-stage method that predicts heat maps for the line segments and junctions using two CNNs and combines junctions and lines by applying several postprocessing steps. To train their method, they created a large wireframe benchmark dataset. Zhou et al. [10] designed an end-to-end trainable L-CNN that directly predicts vectorized wireframes. The L-CNN consists of a stacked hourglass network as the feature extraction backbone, a heat-map-based junction proposal module, a line-sampling module that generates line candidates based on the predicted junctions, and a line verification module, for which the line of interest (LoI) pooling layer is utilized, which compares line segments with corresponding positions in the feature maps of the backbone. The holistically-attracted wireframe parser (HAWP) [11] was built on the L-CNN and introduced a novel line segment reparameterization by using a holistic attraction field map that assigns each pixel to its closest line segment. Lin et al. [12] proposed in their deep Hough transform line priors method to combine line priors with deep learning by incorporating a trainable Hough transform block into a deep network and performing filtering in the Hough domain with local convolutions. For the application of line detection on the Wireframe datasets, they used the L-CNN [10] and the HAWP [11] as backbones and replaced the hourglass blocks with their Hough transform blocks.

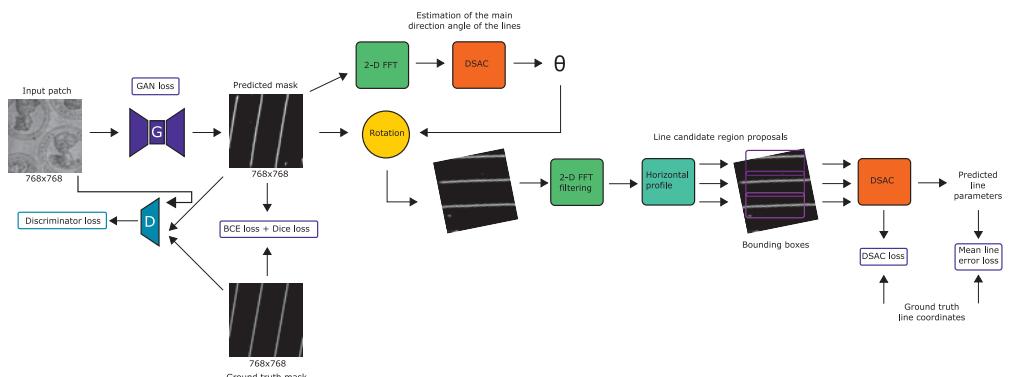
For the application of semantic lines or horizon detection in natural scenes, Lee et al. [13] proposed the VGG16-based semantic line network (SLNet) with line pooling layers, which combines line detection as a multitask loss of classification and regression. The deep Hough transform method by Zhao et al. [14] incorporates the Hough transform into a one-shot end-to-end learning pipeline by using a CNN encoder with feature pyramids for feature extraction and performing the line detection in Hough space. Nguyen et al. [15] transferred the ideas from object detection to design the LS-Net for power line detection that uses a CNN with two heads: one for classification and the other for line regression. Brachmann et al. [16] combined neural guidance with differentiable RANSAC (DSAC) [17] for horizon line estimation.

### 2.3. Contour Detection Using Generative Adversarial Networks

Another related group to our chain line segmentation method consists of contour detection methods using generative adversarial networks (GANs), as the chain lines and contours have a similar shape, and hence, both are sparse segmentation tasks. Contour detection datasets usually contain multiple ground truth annotations per image by different annotators, since the amount of annotated lines differs between the annotators depending on the subjective decision of the individual annotator whether a contour is important enough to be drawn. ContourGAN [18] uses a conditional GAN with a VGG16-based generator network for contour detection in natural images. The adversarial loss is combined with a binary cross-entropy content loss for which the set of ground truth contour images is linearly merged into a single ground truth image. Art2Contour [19] utilizes a conditional GAN with a ResNet-based generator network for salient contour detection in prints and paintings. Art2Contour is trained with a combined loss of the cGAN loss and a task loss consisting of multiple regression terms, which separately treat the single ground truth images. Our method was based on the network architecture used by Art2Contour, but we introduced a novel multitask loss to simultaneously learn line segmentation and line parameterization.

## 3. Method

Our proposed method for the segmentation and detection of chain lines in transmitted light images of historical prints is illustrated in Figure 2. In this section, we introduce the network architecture, the end-to-end trainable pipeline, the loss functions, and inference.



**Figure 2.** ChainLineNet: End-to-end trainable segmentation and parameterization of chain lines using a conditional generative adversarial network-based approach. The generator network is trained using a multitask loss consisting of the segmentation task and the line parameterization task. We propose a fully differentiable pipeline for line coordinates' estimation that is composed of line segmentation, primary line orientation prediction, horizontal alignment of the lines, 2D Fourier filtering, line region proposals, and line fitting with differentiable sample consensus (DSAC) [17]. Detail transmitted light image (input patch): Unknown, *Compilation sheet with round portraits*, Woodcut, Kupferstichkabinett, Staatliche Museen zu Berlin, 44-1884; captured by Thomas Klinke; all rights reserved by the respective museum.

### 3.1. Chain Line Segmentation Network Architecture

Our chain line segmentation network is a conditional generative adversarial network (cGAN) [20] consisting of a generator and discriminator network. Our generator network is the ResNet-based [21] encoder–decoder architecture that was introduced for style transfer [22], having ResNet blocks in the bottleneck, and in contrast to UNet [7], it does not have skip connections between the encoder and decoder [19,23]. As the discriminator network, we used a regular global GAN that has been shown to be effective for contour detection [19,23].

### 3.2. End-to-End Training of Line Segmentation and Parameterization

We jointly trained the generator network for the tasks of line segmentation and line parameterization in an end-to-end fashion by only using differentiable modules and functions inspired by known operator learning [24], while the discriminator network only evaluates the segmentation output against the ground truth segmentation mask.

In generative adversarial networks (GANs), the generator network and the discriminator network are alternately optimized. The generator  $G$  is fed with a random noise vector  $z$  to generate the output image  $y$ , while the discriminator  $D$  is trained to distinguish real images from fake images. In the case of conditional GANs, the output of the generator  $y$  is additionally conditioned to an input, e.g., an image  $x$ . Thus, the generator is trained to generate realistic-looking images that are directly related to the input images. The objective function of cGAN is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{cGAN}}(x, y, z) = & \min_G \max_D \mathbb{E}_{x,y}[\log D(x, y)] \\ & + \mathbb{E}_{x,z}[\log (1 - D(x, G(x, z)))] . \end{aligned} \quad (1)$$

The cGAN principle can be directly applied to the line segmentation task. The generator  $G$  learns to produce precise line segmentation masks  $y \in \mathbb{R}^{s_1 \times s_2}$  for the input artwork images  $x \in \mathbb{R}^{s_1 \times s_2}$ , encouraged by the discriminator  $D$ , which learns to detect those fake ones. The cGAN loss is generally combined with a task loss. We extended this approach by also including the line coordinates' estimation process for the generator task loss:

$$\mathcal{L}_G(x, y, g, h_1, \dots, h_m, p, q) = \mathcal{L}_{\text{cGAN}}(x, y) + \lambda_0 \mathcal{L}_{\text{Task}}(y, g, h_1, \dots, h_m, p, q) , \quad (2)$$

where  $g \in \mathbb{R}^{s_1 \times s_2}$  is the ground truth segmentation mask,  $\{h_1, \dots, h_m\}$  the line hypotheses sampled for DSAC,  $p \in \mathbb{R}^{M \times 4}$  the predicted line coordinates, and  $q \in \mathbb{R}^{N \times 4}$  the ground truth lines coordinates with  $\{x_0^i, y_0^i, x_1^i, y_1^i\}$  being the start and end points of the lines. Our multitask loss is defined as the weighted sum of the line segmentation task and line parameterization task:

$$\begin{aligned} \mathcal{L}_{\text{Task}}(y, g, h_1, \dots, h_m, p, q) = & \lambda_{\text{BCE}} \mathcal{L}_{\text{BCE}}(y, g) + \lambda_{\text{DICE}} \mathcal{L}_{\text{DICE}}(y, g) \\ & + \lambda_{\text{DSAC}} \mathcal{L}_{\text{DSAC}}(h_1, \dots, h_m, q) + \lambda_{\text{MLE}} \mathcal{L}_{\text{MLE}}(p, q) , \end{aligned} \quad (3)$$

where  $\lambda_{\text{BCE}}, \lambda_{\text{DICE}}$  are the weights for the binary cross-entropy loss (BCE) and Dice loss (DICE) for the segmentation task and  $\lambda_{\text{DSAC}}, \lambda_{\text{MLE}}$  are the weights for the DSAC loss [17] and the mean line distance error loss (MLE) for the line parameterization task.

### 3.3. Line Parameterization Pipeline and Line Loss Functions

The prediction of the line parameters is subdivided into the parts of line segmentation, prediction of the main line orientation to horizontally align the lines, 2D Fourier filtering, line region proposals, and line fitting with differentiable sample consensus (DSAC) [17], as illustrated in Figure 2. As chain lines are nearly parallel to each other and have similar distances between them, we used the 2D fast Fourier transform (FFT) to find the main orientation of the lines in the images. The 2D Fourier representation of the segmentation mask shows the response to the dominant direction of the lines.

As can be seen in the centered 2D Fourier magnitude image in Figure 3, there is one line in the center with an orientation orthogonal to that of the chain lines in the image domain. Hence, we extracted the  $k = 500$  points with maximal intensity in the centered magnitude image and fit a line through them using DSAC [17]. Then, we computed the polar angle of the line  $\theta_{ft}$  and determined the rotation angle  $\theta_{rot}$  to align the lines horizontally by:

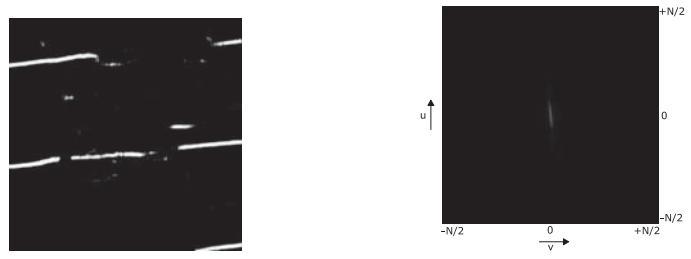
$$\theta_{rot} = \begin{cases} 90^\circ - \|\theta_{ft}\| & \theta_{ft} < 0, \\ 90^\circ + \|\theta_{ft}\| & \text{otherwise} \end{cases} \quad (4)$$

Next, we rotated the predicted segmentation masks, the ground truth segmentation masks, and the ground truth line coordinates; see Figure 4. The segmentation mask can show some line segments of different orientations, for instance due to watermarks, as these have the same intensity as chain lines in the transmitted light images. To reduce nonhorizontal line segments, we applied a vertical filter  $H(u, v)$  in the Fourier domain to the rotated segmentation masks  $F(u, v)$  with  $u, v \in \{-N/2, N/2\}$ :

$$G(u, v) = F(u, v)H(u, v), H(u, v) = \begin{cases} 1 & \|v\| < \tau, \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

As convolution with a filter kernel in the time domain is elementwise matrix multiplication in the Fourier domain, we can simply multiply the 2D Fourier image with a matrix that has only zero elements except for a vertical band of width  $2\tau$  with  $\tau = 10$  pixels at the center.

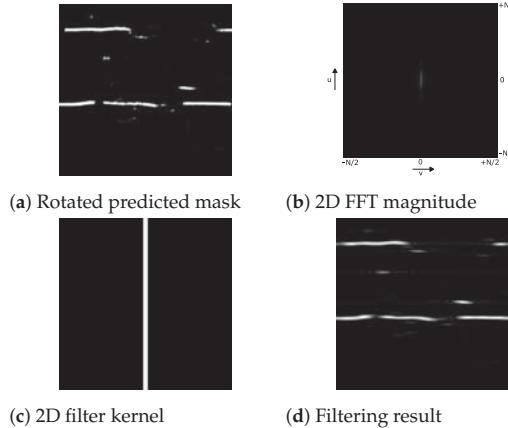
To determine the number of lines and their rough positions, we computed the horizontal profile by summing up all intensity values of the filtered segmentation mask along the  $x$ -direction (see Figure 5). All segmented lines correspond to peaks in the profile. We applied 1D max-pooling to the profile to obtain all local maxima. To filter out all local maxima that most likely do not belong to the horizontal lines, we applied intensity and spatial distance thresholding. As prior knowledge, we considered that chain lines have approximately the same distances; hence, we first selected the peaks that have a distance of at least 0.75 of the maximal distance of all peaks and then refined the selection by keeping only those that have at least 0.6 of the maximal distance of the selected peaks.



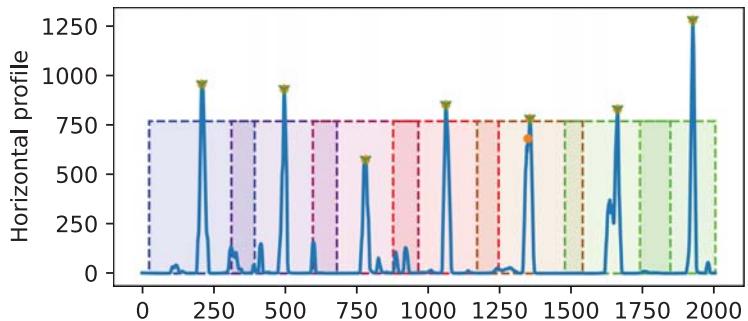
(a) Predicted mask

(b) 2D fast Fourier transform magnitude

**Figure 3.** The 2D fast Fourier transform of the predicted segmentation mask in (b) shows a centered line whose orientation is orthogonal to the dominant orientation of the line segments in (a) the predicted segmentation mask.



**Figure 4.** Two-dimensional filtering in Fourier domain to reduce nonhorizontal lines. In (a), the horizontally aligned predicted segmentation mask, in (b), its 2D FFT magnitude, in (c), the 2D filter kernel, and in (d), the filtering result of the rotated predicted segmentation mask is shown.



**Figure 5.** Selection of peaks in the horizontal profile for one example image (length of 2000 pixels) with 7 chain lines from the validation set. The peaks are marked with an orange dot, and the final selection of peaks after distance thresholding are additionally marked with a green cross. Then, a bounding box is placed at the center of each selected peak.

A horizontally oriented bounding box  $[0, W, y_i - D_{\text{thresh}}, y_i + D_{\text{thresh}}]$  is defined for each of the refined peak positions  $y_i$  using the previously computed threshold  $D_{\text{thresh}}$  as the length to both sides. In the case that no peak position can be found or can be selected, we defined one bounding box for the entire image.

Next, we extracted for each bounding box the maximal  $k_p$  points within the bounding box region of the segmentation mask to use them for line fitting with DSAC. DSAC [17] formulates the hard hypothesis selection of RANSAC as a probabilistic process that allows end-to-end learning. The application of DSAC for line fitting (implementation by Brachmann et al.: <https://github.com/vislearn/DSACLine> (accessed on 14 July 2021)) consists of the following steps:

1. Line hypothesis sampling: Based on the predicted point coordinates  $z$ ,  $m$  line hypotheses  $\{h_1, \dots, h_m\}$  are randomly sampled by choosing for each hypothesis two points of the point set. Each hypothesis predicts an estimate for the line parameters, the slope  $a$  and intercept  $b$  of the line equation  $y = ax + b$ ;

2. Hypothesis selection: A scoring function  $s(h)$  computes a score for each hypothesis based on the soft inlier count. The hypothesis  $h_j$  is selected according to the softmax probabilistic distribution  $P(j; z) = \frac{\exp(s(h_j))}{\sum_k \exp(s(h_k))}$ ;
3. Hypothesis refinement: The hypothesis is refined by using the weighted Deming regression for line fitting [25], which is a special case of the total least-squares that accounts for errors in the observations in both the  $x$ - and  $y$ -direction, for which we used the soft inlier scores as the weights.

The DSAC loss function, which we incorporated into our task loss function, is defined as:

$$\mathcal{L}_{\text{DSAC}}(h_1, \dots, h_m, \mathbf{q}) = \sum_k \left( \frac{\exp(s(h_j))}{\sum_k \exp(s(h_k))} \|\mathbf{p}(h_k) - \mathbf{q}\|_2 \right), \quad (6)$$

where  $\mathbf{p}(h_k)$  refers to the predicted start and end points for the line hypothesis  $h_k$  and  $\mathbf{q}$  refers to the ground truth start and end points. The start and end points of the lines are determined as the intersection with the image borders.

Since we applied DSAC to each bounding box region separately and the bounding box positions are determined automatically based on the segmentation output of the network, we needed to assign one ground truth line to each bounding box. We distinguish three cases: (1) If there is only one ground truth line inside the bounding box region, this one is selected. (2) If the region contains multiple ground truth lines, we chose the longest line. (3) Lastly, if there is no ground truth line inside the region, we selected the line with the minimal distance of its start and end points to the borders of the region.

The DSAC loss minimizes the distance of the predicted lines to the closest ground truth lines; however, if too few bounding boxes are predicted, some ground truth lines will not be included. To account for these false negatives, we defined a second line loss term, the MLE loss, that picks for each ground truth line the closest predicted line of the best hypothesis  $h_j$  and computes the mean error:

$$\mathcal{L}_{\text{MLE}}(\mathbf{p}, \mathbf{q}) = \frac{1}{N} \sum_i \min(\mathbf{D}_i), \quad \mathbf{D} = \text{cdist}(\mathbf{p}, \mathbf{q}), \quad (7)$$

where  $\mathbf{D} \in \mathbb{R}^{N \times M}$  is the Euclidean distance between each pair of the two collections of row vectors of  $\mathbf{p} \in \mathbb{R}^{M \times 4}$ ,  $\mathbf{q} \in \mathbb{R}^{N \times 4}$ , and  $\mathbf{D}_i$  is the  $i^{\text{th}}$  row of the distance matrix.

### 3.4. Inference of Chain Line Segmentation and Parameterization Network

Since the network architecture is fully convolutional, the complete images are fed to the GAN and are processed in the same manner as for training, resulting in the line predictions of the rotated image. Hence, to obtain the final line coordinate predictions of the original image, the inverse rotation is applied to the predicted lines.

## 4. Experiments and Results

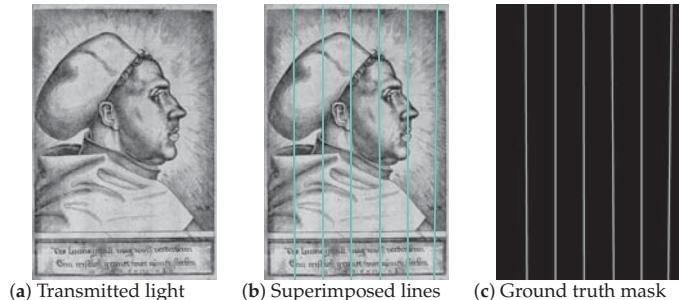
In this section, we describe our dataset for chain line detection in historical prints, we evaluate the performance of our method for line segmentation and line parameterization, and compare it to the state-of-the-art methods and to manual line measurements.

### 4.1. Chain Line Dataset

The dataset consists of high-resolution grayscale transmitted light images of prints from the 16th Century, including portraits of Martin Luther and contemporaries. For our dataset, we selected in total 95 images in which the chain lines were recognizable by the human eye. All images contain chain lines that are either horizontally or vertically distributed at approximately the same distances.

We manually annotated the chain lines in the images by selecting two points on each line and fitted a straight line through them, as illustrated in Figure 6a,b. We used the  $x$  and  $y$  coordinates of the start and end points, as well as the corresponding mask images

(Figure 6c) that contain the segmented ground truth lines as labels for training, validation, and testing.



**Figure 6.** Illustration of the line annotation (a) in the transmitted light images of historical prints by (b) selecting start and end points of the lines and (c) computing the corresponding mask images. Image: (a) Daniel Hopfer, *Martin Luther with the doctor's cap*, Etching, Germanisches Nationalmuseum Nürnberg, K722; captured by Thomas Klinke; all rights reserved by the respective museum.

The sharp edges of the annotated lines in the mask images are smoothed by applying a Gaussian filter with a standard deviation of 3. The images are divided into 35 images for training, 12 images for validation, and 48 images for testing. The images were acquired at a very high resolution with image sizes up to  $5000 \times 6500$  pixels. Since chain lines are very fine structures that are difficult to detect, the highest possible image resolution is recommended, but is limited by hardware constraints. To be able to feed the entire image at once for inference using one Nvidia Titan XP GPU (NVIDIA Corporation, Santa Clara, CA, USA), we scaled all images to the maximal length of 2000 pixels, which is sufficient for the chain line detection task. To train the neural network, we split the scaled images of the training and validation set into image patches of size  $768 \times 768$  pixels with an overlap stride of 384. The image patches contain between one and five lines per patch. Patches that do not contain any line were excluded from training. Further, we applied offline data augmentation (see below) to double the number of training and validation patches, resulting in 1150 training and 370 validation patches.

#### 4.2. Implementation Details

Our method was implemented using the PyTorch framework, and the end-to-end training and inference both ran completely on the GPU. The generator network (9 ResNet blocks) and the discriminator network were trained from scratch for 100 epochs with early stopping by using the Adam optimizer, a learning rate of  $\eta = 0.0002$  with linear decay to 0 starting at Epoch 50, momentum (0.5, 0.999), a batch size of 2,  $\lambda_0 = 1000$  [19],  $\lambda_{BCE} = 0.5$ ,  $\lambda_{DICE} = 0.5$ ,  $\lambda_{DSAC} = 0.5$ , and  $\lambda_{MLE} = 0.5$ . For DSAC,  $m = 64$  hypotheses are sampled based on  $k_p = 500$  points from each bounding box per patch or  $k_p = 1300$  points from each bounding box per image.

Prior to training, we augmented our training and validation set in an offline manner with rotated images, i.e., rotations by 90 degrees were applied to produce the same number of vertical and horizontal lines. During training, we applied online data augmentation (color jittering, blurring, horizontal and vertical flipping, and rotation with angles uniformly sampled in the range of  $(-20, 20)$  degrees) only to the training set, and not to the validation set.

#### 4.3. Evaluation of Line Segmentation

In this section, we compare different architectures for the task of chain line segmentation using pixelwise precision, recall, and the Dice coefficient (i.e., pixelwise F1-score) of the predicted segmentation results and ground truth segmentations. To compute the

metrics, we applied a threshold of 0.5 to binarize the segmentation masks. For this experiment, all networks were trained only for the segmentation task (i.e.,  $\lambda_{BCE} = \lambda_{DICE} = 0.5$ ,  $\lambda_{DSAC} = \lambda_{MLE} = 0$ ). We compared the UNet (with feature dimension  $F = 16$ ; 1,942,289 parameters, and  $F = 64$ ; 31,036,481 parameters) and the ResNet-based encoder–decoder architecture ( $F = 64$ ; 11,370,881 parameters) alone and plugged into the generative adversarial training as generator networks. As summarized in Table 1 for the validation set, all network architectures achieve higher recall than precision. Precision is highest for the small UNet-GAN and recall for the ResNet-GAN, directly followed by the ResNet encoder–decoder (ResNet-E-D). The Dice coefficient, which combines the pixelwise precision and recall into one measure, is also highest for the ResNet-GAN and second best for the ResNet encoder–decoder. Concerning the Dice coefficient, UNet seems not to profit from adversarial training in our specific case. Based on these observations, we chose the ResNet-GAN architecture for our end-to-end trainable line segmentation and detection method.

**Table 1.** Evaluation of pixelwise precision, recall, and the Dice coefficient for chain line segmentation of the validation set with 12 images. Best scores are highlighted in bold.

Method	Precision	Recall	Dice Coefficient
UNet ( $F = 16$ )	0.4046	0.5070	0.4464
UNet ( $F = 64$ )	0.3958	0.5034	0.4392
UNet-GAN ( $F = 16$ )	<b>0.4283</b>	0.4787	0.4437
UNet-GAN ( $F = 64$ )	0.3829	0.4591	0.4108
ResNet-E-D ( $F = 64$ )	0.3855	0.5935	0.4628
ResNet-GAN ( $F = 64$ )	0.3920	<b>0.6001</b>	<b>0.4696</b>

#### 4.4. Evaluation of Line Detection and Parameterization

For the evaluation of line detection and parameterization, we compared the number of predicted lines using precision, recall, and the  $F_1$  score. Therefore, we counted the number of true positives, false positives, and false negatives based on a pixel distance threshold of 50 by computing the distance between the start and end point of the predicted lines and ground truth lines that were manually annotated on the digital images. As a metric, we computed the mean pixel differences of chain line positions w.r.t. the ground truth line coordinates only for the true positive lines. Furthermore, we compared the automatically computed chain line distance intervals with the manual measurement of an art technologist, who has measured the chain line distance intervals directly on the physical paper during his art technological examination. To convert the predicted pixel distance intervals into distance intervals in millimeters such that these can be directly compared to the physical measurements, we scaled the images based on the manually measured width of the artwork. For the chain line distance comparison, we only considered images in which both the number of true positive lines and the total number of detected lines differs only at most by about 2 lines from the number of reference lines by the art technologist. We used cross-correlation to automatically find the best position to arrange the two distance intervals as they can be shifted against each other if one or two lines are not detected. Then, we computed the mean absolute difference of the overlap of both intervals.

##### 4.4.1. Ablation Study

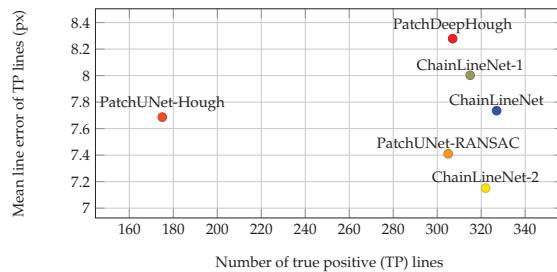
We evaluated the influence of our ChainLineNet using all task loss terms in contrast to setting individual terms to zero. First, we compare the line detection results in Table 2 for the test set. By using our novel multitask loss consisting of the segmentation losses (BCE+DICE) and the line parameterization losses (DSAC+MLE), we achieved a gain in the  $F_1$  score of about 1% in comparison to training the network only for the segmentation task (ChainLineNet-2) and of about 2% in comparison to the end-to-end training only by using the BCE+DICE+DSAC losses (ChainLineNet-1). The DSAC loss alone does not consider false negatives, hence resulting in a lower recall.

**Table 2.** Evaluation of precision, recall, and the  $F_1$  score of chain line detection for the test set with 48 images. The number of true positives (TP), false positives (FP), and false negatives (FN) are determined based on a distance threshold of 50 pixels between the predicted and ground truth lines. Best scores are highlighted in bold.

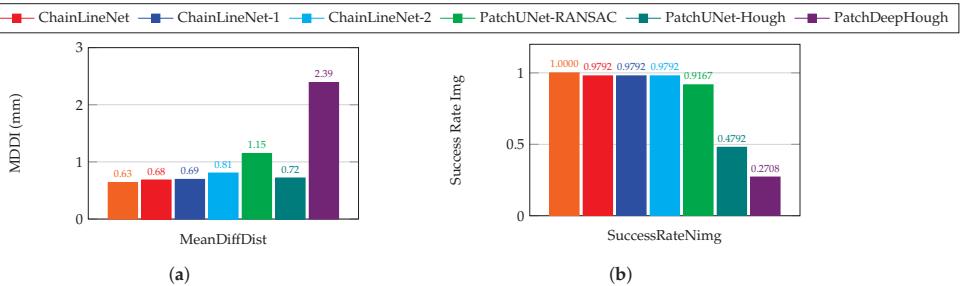
	Number of Lines	TP	FP	FN	Precision (%)	Recall (%)	$F_1$ Score (%)
Ground truth (manually annotated)	342	342	0	0	100.00	100.00	100.00
Reference (manually measured)	339	339	0	3	<b>100.00</b>	<b>99.12</b>	<b>99.56</b>
PatchDeepHough	528	307	221	35	58.14	89.77	70.57
PatchUNet-Hough	228	175	53	162	76.75	51.93	61.95
PatchUNet-RANSAC	325	305	20	32	93.85	90.50	92.15
ChainLineNet-1 (BCE+DICE+DSAC)	323	315	8	27	97.52	92.11	94.74
ChainLineNet-2 (BCE+DICE)	330	322	8	20	97.58	94.15	95.83
ChainLineNet (BCE+DICE+DSAC+MLE)	333	327	6	15	<b>98.20</b>	<b>95.61</b>	<b>96.89</b>

Secondly, we compared the difference of the line positions between the predicted and ground truth lines in Figure 7 for the test set. The line error was only calculated for true positives. The mean line error of true positives lies between 7 and 8 pixels with the lowest error for ChainLineNet-2 (only segmentation), followed by ChainLineNet (all losses) and ChainLineNet-1 (segmentation + DSAC). However, the results are very close, and the number of true positives of the ChainLineNet is a bit higher, which could be a reason for the slightly higher pixel error of almost 0.6 in comparison to ChainLineNet-2.

Lastly, we compare in Figure 8, for the test set, the distance intervals for the images that contain a suitable number of lines with the reference distance measurements. For this comparison (see Figure 8b), only one image was excluded, giving a success rate of about 98% for all versions of ChainLineNet. The mean difference of the distance intervals (Figure 8a) is below 1 mm for all three variants, whereas ChainLineNet (all losses) achieves the best result, directly followed by ChainLineNet-1 (with DSAC) and ChainLineNet-2 (only segmentation) being a bit inferior.



**Figure 7.** Comparison of the mean pixel line error between the true positive predicted line coordinates and the ground truth line coordinates for the test set. Our ChainLineNet (complete task loss) is compared to the end-to-end training with the task losses BCE+DICE+DSAC (ChainLineNet-1), to the training using only the segmentation task losses BCE+DICE (ChainLineNet-2), and to the competing methods.



**Figure 8.** Comparison of (a) the mean difference of distance intervals (MDDI) between the predicted distances and the reference distances (manually measured by an art technologist) and (b) the success rate of images for which the distance intervals were compared. Our ChainLineNet (complete task loss) is compared to the end-to-end training with the task losses BCE+DICE+DSAC (ChainLineNet-1), to the training using only the segmentation task losses BCE+DICE (ChainLineNet-2), to the competing methods, and to the ground truth on the test set.

#### 4.4.2. Comparison to the State-of-the-Art

In this section, we measure the performance of our ChainLineNet compared to competing methods. We retrained the UNet architecture ( $F = 16$ ) of our previous work [6], which was implemented in TensorFlow, for our renewed historical print dataset for 30 epochs using a learning rate of  $\eta = 0.0001$  and a batch size of 5. During inference, the UNet was executed patchwise, and two postprocessing methods were applied to the reassembled segmentation output [6], which we refer to as PatchUNet-RANSAC and PatchUNet-Hough. Secondly, we trained the deep Hough transform line prior method [12] for our line detection task, which we abbreviate as PatchDeepHough. The method was originally developed for wireframe detection; thus, some modifications were necessary to make it applicable to our task. We used their offline data augmentation, which quadrupled the number of training patches, and trained the network from scratch for 50 epochs with early stopping using a learning rate of  $\eta = 0.0004$  and a batch size of 4. Due to the high complexity of the voting matrix needed for the Hough transform, we were not able to increase the input size of the network for inference, such that we used the default setting of  $512 \times 512$  and applied the method patchwise. We added the following postprocessing steps to filter, merge, and extend line segments to full lines: First, we computed the dominant orientation of the line segments, i.e., horizontal or vertical. Then, we excluded all line segments with the opposite orientation and whose Hough score was below 0.7. For the remaining line segments, we walked along the perpendicular direction of the line segments and grouped the segments within a neighborhood of 20 pixels. For each group, we used linear least-squares regression to fit a line through the start and end points of the line segments. In the case of a vertical main orientation of the lines, we switched the  $x$  and  $y$  coordinates for line fitting to obtain more accurate results.

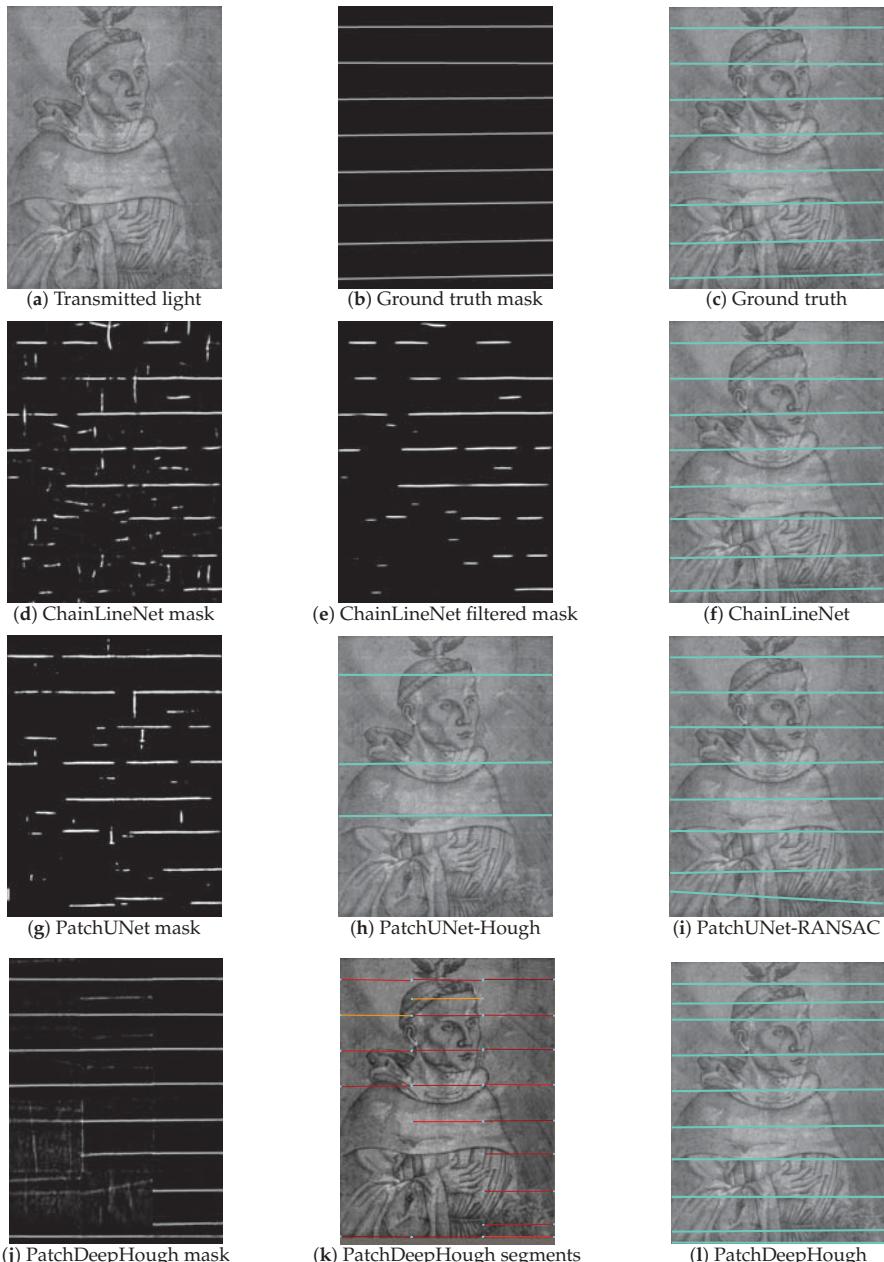
The quantitative results for line detection and parameterization for the test set consisting of 48 images and in total 342 correct lines are summarized in Table 2 for precision, recall, and the  $F_1$ -score. ChainLineNet outperformed all machine learning methods with an  $F_1$ -score of 96.9 %, precision of 98.2 %, and recall of 95.6 %, being close to manual measurements, which obtain an  $F_1$ -score of 99.6 %. In comparison to PatchUNet-RANSAC, which also performs quite well, we achieved an absolute gain of about 4 % in the  $F_1$ -score. PatchDeepHough detects too many false positive lines; thus, it only achieved poor precision and a clearly lower  $F_1$ -score of 70.6 %. PatchUNet-Hough detects distinctively less correct lines, resulting in a low recall and the lowest  $F_1$ -score of 62 %.

The comparison for the pixel mean line error of true positive lines, depicted in Figure 7, shows that all methods predict the line coordinates comparably accurately with an error between 7.2 and 8.3 pixels. The result of ChainLineNet with 327 out of 342 correct lines is the most reliable, as most lines were used to compute the mean line error.

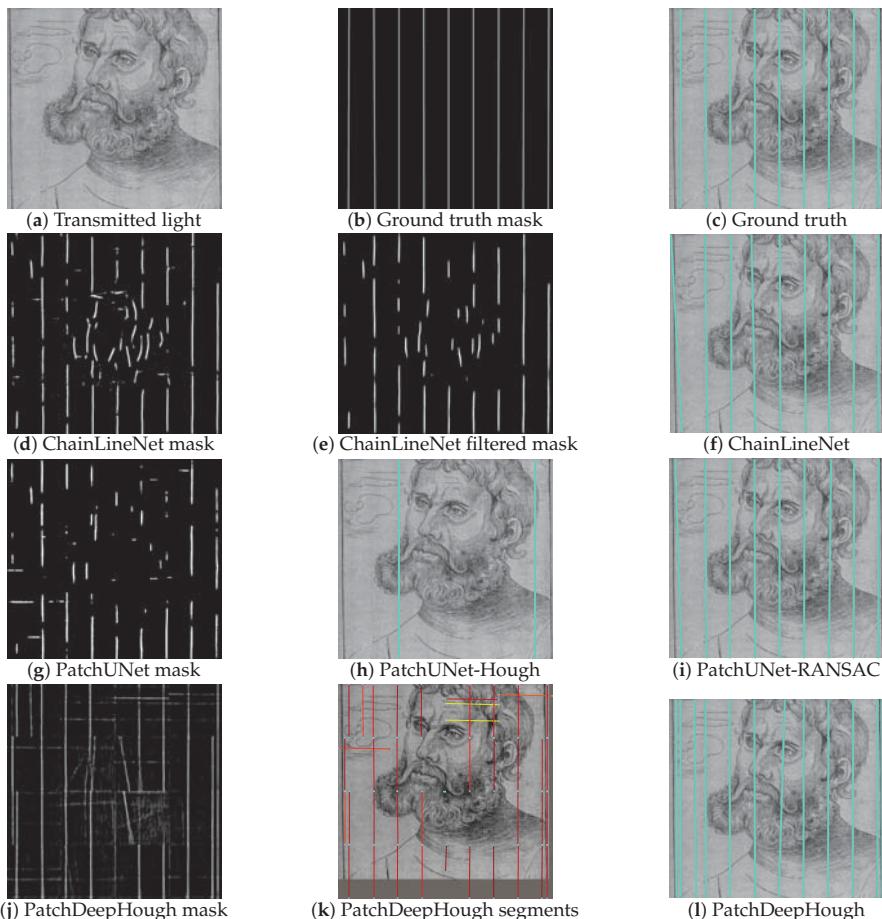
Next, we compare the chain line distance intervals to the reference measurements in Figure 8. The chain line distance intervals computed using ChainLineNet for 47 out of 78 test images only differ by 0.68 mm from the reference intervals, which is an excellent result, when we consider that the comparison of the manually annotated ground truth lines and the reference lines differs by 0.63 mm. Plausible reasons for the measurement inaccuracies are the conversion of the images of the artworks to millimeters, the fact that the location where the line distances are measured can differ between manual and digital measurements, and that chain lines are approximated as straight lines. The other tested machine learning methods show less precision for the distance interval computation. PatchUNet-Hough has a slightly higher mean difference, but only less than half of the images are suitable for the comparison (see Figure 8b). PatchUNet-RANSAC has a slightly lower success rate than ChainLineNet with their mean difference lying just above 1 mm. PatchDeepHough performs worst. With only a success rate of 27 % of the images, their mean difference is above 2 mm.

The qualitative results are shown in Figure 9 for one example with horizontal chain lines and in Figure 10 for an example with vertical chain lines. For both figures, the transmitted light image of the artwork, the ground truth segmentation mask, and the ground truth lines superimposed on the artwork are depicted in the first row. Figures 9d and 10d show the raw segmentation outputs of the ChainLineNet that contain line segments and noise. The noise is reduced in Figures 9e and 10e by 2D Fourier filtering. Here, the filtered mask images are binarized for visualization, because only the points with maximal intensity are selected for DSAC. In Figures 9f and 10f, the final line parameterization results of ChainLineNet are shown, which are in high accordance with the ground truth lines. Figures 9g and 10g show the binarized segmentation output of PatchUNet that is also composed of line segments and noise. Two different postprocessing approaches are applied to the PatchUNet output. PatchUNet-Hough (Figures 9h and 10h) detects clearly fewer lines than PatchUNet-RANSAC (Figures 9i and 10i). The grayscale heat map of PatchDeepHough in Figures 9j and 10j shows many clear lines, but also areas of uncertainty. Due to the patchwise application, line segments are separately fitted in each patch (Figures 9k and 10k), where the Hough voting score is indicated by the line segment color ranging from low (blue) to high (red). PatchDeepHough predicts clearly too many lines, as can be seen in Figures 9l and 10l. Despite the watermark that is included in the paper structure of Figure 10a, all methods are able to detect chain lines that interfere with the watermark.

Overall, our method achieves excellent performance, but there are some limitations. In the case of bent wires, our method cannot determine the exact chain line, but only an approximation, because we assumed straight lines for our model. Difficult images, where the chain lines are densely covered with ink, the paper is in an abraded condition, or when lines in the border area of the image are only partly depicted, can lead to false positives or false negatives. Under very difficult image conditions, the application of DSAC can lead to inaccurate line predictions, e.g., if a too large bounding box size is determined by our method or the estimated rotation angle is not accurate enough. In these cases, the bounding box might contain line segments or noise that do not belong to the actual line. Difficult cases need to be reviewed by art technologists, but our method achieves a high success rate such that it can greatly support the art technologists in their analysis of the artworks.



**Figure 9.** Qualitative results of the chain line detection for one historical print containing horizontal chain lines. Transmitted light image: Hieronymus Hopfer, *Martin Luther as Augustinian monk with Holy Spirit as a dove*, Etching, British Museum, London, 1845-0809-1486; Photo © Thomas Klinke, courtesy of the Trustees of the British Museum; all rights reserved by the respective museum.



**Figure 10.** Qualitative results of the chain line detection for one historical print containing vertical chain lines and a watermark. Transmitted light image (detail): After Lucas Cranach the Elder, *Martin Luther as Junker Jörg*, Collotype, Kunstsammlungen der Veste Coburg, H.0064; captured by Thomas Klinke; all rights reserved by the respective museum.

## 5. Conclusions

We presented an end-to-end trainable deep learning method for chain line segmentation and parameterization in historical prints. In the experiments, we showed that our ChainLineNet achieves the best visual and quantitative chain line detection results for our historical print dataset. Moreover, the comparison of the automatically computed chain line distance intervals with the manually measured distance intervals by an art technologist shows a low error of less than 0.7 mm. The high accuracy and reliability of our method give the opportunity to automatically compare the chain line distances of a larger number of historical prints in order to draw conclusions about the origin of the papers. Thus, our automatic deep-learning-based method can be very beneficial to support the art historical and technological research of museums and libraries. Future work could build on the automatic chain line detection and distance computation to extract chain line distance patterns and perform a similarity search to identify moldmates.

**Author Contributions:** Conceptualization, A.S., T.K., A.M. and V.C.; methodology, A.S.; software, A.S.; validation, A.S.; formal analysis, A.S.; investigation, A.S.; data curation, A.S. and T.K.; writing—original draft preparation, A.S.; writing—review and editing, A.S., T.K., A.M. and V.C.; visualization, A.S.; supervision, A.M. and V.C. All authors read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Leibniz Society Grant Number SAW-2018-GNM-3-KKLB.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data used for the development and evaluation of the method was collected within the research project “Critical catalogue of Luther portraits (1519–1530)” by the Germanisches Nationalmuseum Nürnberg, FAU Erlangen-Nürnberg and TH Köln, in which A.S., T.K., A.M., and V.C. are involved. The photographs of the historical prints were captured by T.K., all rights reserved by the respective museum/library. The image data of the research project can soon be viewed online at the Cranach Digital Archive (<https://lucascranach.org/>).

**Acknowledgments:** Thanks to Daniel Hess, Germanisches Nationalmuseum, Gunnar Heydenreich, Cranach Digital Archive, for providing image data and NVIDIA Corporation for their GPU hardware donation.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Johnson, C.R.; Sethares, W.A.; Ellis, M.H.; Haqqi, S. Hunting for Paper Moldmates Among Rembrandt’s Prints: Chain-line pattern matching. *IEEE Signal Process. Mag.* **2015**, *32*, 28–37. [[CrossRef](#)]
- Hiary, H.; Ng, K. A system for segmenting and extracting paper-based watermark designs. *Int. J. Digit. Libr.* **2007**, *351*–361. [[CrossRef](#)]
- van der Lubbe, J.; Someren, E.; Reinders, M.J. Dating and Authentication of Rembrandt’s Etchings with the Help of Computational Intelligence. In Proceedings of the International Cultural Heritage Informatics Meeting (ICHIM), Milan, Italy, 3–7 September 2001; pp. 485–492
- Atanasius, V. Assessing paper origin and quality through large-scale laid lines density measurements. In Proceedings of the 26th Congress of the International Paper Historians Association, Rome/Verona, Italy, 30 August–6 September 2002; pp. 172–184.
- van Staalduinen, M.; van der Lubbe, J.; Backer, E.; Paclík, P. Paper Retrieval Based on Specific Paper Features: Chain and Laid Lines. In Proceedings of the Multimedia Content Representation, Classification and Security (MRCS) 2006, Istanbul, Turkey, 11–13 September 2006; pp. 346–353. [[CrossRef](#)]
- Biendl, M.; Sindel, A.; Klinke, T.; Maier, A.; Christlein, V. Automatic Chain Line Segmentation in Historical Prints. In Proceedings of the Pattern Recognition, ICPR International Workshops and Challenges, Milan, Italy, 10–15 January 2021; Springer International Publishing: Cham, Switzerland, 2021; pp. 657–665. [[CrossRef](#)]
- Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2015, Munich, Germany, 5–9 October 2015; Volume 9351, pp. 234–241. [[CrossRef](#)]
- Fischler, M.A.; Bolles, R.C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM* **1981**, *24*, 381–395. [[CrossRef](#)]
- Huang, K.; Wang, Y.; Zhou, Z.; Ding, T.; Gao, S.; Ma, Y. Learning to Parse Wireframes in Images of Man-Made Environments. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 626–635. [[CrossRef](#)]
- Zhou, Y.; Qi, H.; Ma, Y. End-to-End Wireframe Parsing. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; [[CrossRef](#)]
- Xue, N.; Wu, T.; Bai, S.; Wang, F.; Xia, G.S.; Zhang, L.; Torr, P.H. Holistically-Attracted Wireframe Parsing. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020. [[CrossRef](#)]
- Lin, Y.; Pintea, S.L.; van Gemert, J.C. Deep Hough-Transform Line Priors. In Proceedings of the European Conference on Computer Vision (ECCV) 2020, Glasgow, UK, 23–28 August 2020; Volume 12367, pp. 323–340. [[CrossRef](#)]
- Lee, J.T.; Kim, H.U.; Lee, C.; Kim, C.S. Semantic Line Detection and Its Applications. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3249–3257. [[CrossRef](#)]
- Zhao, K.; Han, Q.; Zhang, C.B.; Xu, J.; Cheng, M.M. Deep Hough Transform for Semantic Line Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [[CrossRef](#)] [[PubMed](#)]
- Nguyen, V.N.; Jenssen, R.; Roverso, D. LS-Net: Fast single-shot line-segment detector. *Mach. Vis. Appl.* **2020**, *1432*–1769. [[CrossRef](#)]

16. Brachmann, E.; Rother, C. Neural-Guided RANSAC: Learning Where to Sample Model Hypotheses. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 4321–4330. [[CrossRef](#)]
17. Brachmann, E.; Krull, A.; Nowozin, S.; Shotton, J.; Michel, F.; Gumhold, S.; Rother, C. DSAC—Differentiable RANSAC for Camera Localization. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; [[CrossRef](#)]
18. Yang, H.; Li, Y.; Yan, X.; Cao, F. ContourGAN: Image contour detection with generative adversarial network. *Knowl.-Based Syst.* **2019**, *164*, 21–28. [[CrossRef](#)]
19. Sindel, A.; Maier, A.; Christlein, V. Art2Contour: Salient Contour Detection in Artworks Using Generative Adversarial Networks. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 788–792. [[CrossRef](#)]
20. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [[CrossRef](#)]
22. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Volume 9906, pp. 694–711. [[CrossRef](#)]
23. Li, M.; Lin, Z.; Mech, R.; Yumer, E.; Ramanan, D. Photo-Sketching: Inferring Contour Drawings from Images. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1403–1412. [[CrossRef](#)]
24. Maier, A.; Syben, C.; Stimpel, B.; Würfl, T.; Hoffmann, M.; Schebesch, F.; Fu, W.; Mill, L.; Kling, L.; Christiansen, S. Learning with known operators reduces maximum error bounds. *Nat. Mach. Intell.* **2019**, *1*, 2522–5839. [[CrossRef](#)] [[PubMed](#)]
25. Linnet, K. Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clin. Chem.* **1998**, *44*, 1024–1031. [[CrossRef](#)] [[PubMed](#)]



MDPI  
St. Alban-Anlage 66  
4052 Basel  
Switzerland  
Tel. +41 61 683 77 34  
Fax +41 61 302 89 18  
[www.mdpi.com](http://www.mdpi.com)

*Journal of Imaging* Editorial Office  
E-mail: [jimaging@mdpi.com](mailto:jimaging@mdpi.com)  
[www.mdpi.com/journal/jimaging](http://www.mdpi.com/journal/jimaging)





MDPI  
St. Alban-Anlage 66  
4052 Basel  
Switzerland

Tel: +41 61 683 77 34  
Fax: +41 61 302 89 18  
[www.mdpi.com](http://www.mdpi.com)



ISBN 978-3-0365-2226-5