

Identifying AI-Generated Art with Deep Learning

Tommaso Bianco, Giovanna Castellano, Raffaele Scaringi* and Gennaro Vessio

Department of Computer Science, University of Bari Aldo Moro, Bari, Italy

Abstract

Generative AI, mainly through Diffusion Models, has revolutionized art creation, blurring the distinction between human and AI-generated art. This study introduces a novel dataset comprising human-made and AI-generated art and employs Deep Learning models (VGG-19, ResNet-50, ViT) to distinguish between them. We also use eXplainable AI techniques to derive insights. Our results highlight the potential of AI to detect machine-generated content, with implications for art authentication.

Keywords

Computer vision, Deep learning, Digital humanities, Generative AI, Synthetic art

1. Introduction

Art has undergone a profound transformation with the emergence of generative Artificial Intelligence, notably driven by technologies such as Generative Adversarial Networks [1] and the increasingly popular Diffusion Models [2]. These groundbreaking innovations have pushed the boundaries of artistic creation, empowering machines to produce remarkably lifelike images, including paintings, that challenge our conventional notions of human creativity. Indeed, generative AI has exhibited the capability to generate synthetic paintings that closely emulate renowned artists' styles, brushwork, and aesthetics. This level of fidelity in replicating the artistic process blurs the demarcation between traditional, human-crafted art and machine-generated creations.

The distinction between genuine human-made art and its synthetic counterparts carries extensive implications, influencing aspects such as art authentication, valuation, and preservation while igniting debates concerning technology's role in the creative process. While conventional methods of art connoisseurship traditionally relied on expert human judgment, the rapid evolution of Deep Learning models and the availability of extensive art datasets present new avenues for addressing this challenge. One intriguing approach to detecting instances generated by machines, in fact, involves leveraging the capabilities of machines themselves. This concept is rooted in the idea that the same AI technologies responsible for creating synthetic content can also be employed for their detection and differentiation from authentic human-made counterparts. Deep Learning and Computer Vision algorithms can also be trained on large datasets containing authentic and AI-generated examples. These models can learn to identify

CREAI 2023: 2nd Workshop on Artificial Intelligence and Creativity, November 6–9, 2023, Rome, Italy

*Corresponding author.

✉ t.bianco5@studenti.uniba.it (T. Bianco); giovanna.castellano@uniba.it (G. Castellano); raffaele.scaringi@uniba.it (R. Scaringi); gennaro.vessio@uniba.it (G. Vessio)

ORCID 0000–0002–6489–8628 (G. Castellano); 0000–0001–7512–7661 (R. Scaringi); 0000–0002–0883–2691 (G. Vessio)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

subtle patterns, inconsistencies, or artifacts that may indicate machine generation. By utilizing AI-powered classification models, we can automate the process of detecting machine-generated content, achieving both efficiency and accuracy.

Given these considerations, this paper distinguishes between authentic and synthetically generated artworks. This effort begins with creating a novel dataset and proceeds to conduct experiments that evaluate well-established Deep Learning algorithms, including VGG-19 [3], ResNet-50 [4], and Vision Transformers (ViT) [5]. Furthermore, we complement our quantitative analysis with qualitative examinations using eXplainable AI frameworks, particularly Grad-CAM [6]. This approach can help visually justify why a given algorithm classifies an image as real or synthetic, shedding light on the most discriminative image regions.

The rest of this paper is structured as follows. Section 2 reviews the existing literature on recognizing synthetic images. Section 3 introduces our proposed dataset. Section 4 details the classification method proposed for this task. Section 5 provides insights into the experiments conducted in this research. Finally, Sec. 6 concludes the article while highlighting potential directions for future research efforts.

2. Related Work

In recent years, there has been a significant surge in interest in Deep Learning-based models, particularly in image analysis. Notably, the emergence of Diffusion Models, as exemplified in the review paper by Croitoru et al. [2], has yielded remarkable outcomes in generating high-fidelity images full of authentic details. However, generative models are often characterized by discernible idiosyncratic patterns, which can be exploited to ascertain an image’s genuineness. Despite concerted efforts to mitigate the presence of such patterns within Diffusion Models, it is imperative to acknowledge that they are not entirely free of such distinctive traits. Specifically, research has underscored the significance of features such as color band inconsistency [7] and the paucity of variation in color intensity [8] in identifying synthetic images.

Consequently, various Computer Vision techniques have been harnessed to facilitate the automated detection of the factors mentioned above. Guo et al. [9] and Guarnera et al. [10] have introduced hierarchical fine-grained classification methodologies explicitly designed to discriminate between forged or synthetic images. Central to their approach is the meticulous curation of a training dataset, wherein the hierarchical framework necessitates the comprehensive incorporation of diverse forgery techniques. This endeavor, however, poses a challenge, particularly in scenarios characterized by limited diversity within the available training data. Addressing the recognition of semantic and stylistic features in images, Amoroso et al. [11] have tackled the issue of discerning synthetic images that exhibit heightened distinctiveness within the stylistic domain. Nonetheless, the practical implementation of semantic-style disentanglement presents notable challenges, primarily requiring the development of bespoke training datasets tailored explicitly to this specific objective. Hamid et al. [12] proposed a framework for fake image detection based on ResNet, achieving good results on a dataset containing real and fake images generated by GANs.

In the digital humanities domain, identifying forgeries, often manifesting as synthetic images, holds paramount significance for art experts. This task tests whether a given artwork is an

authentic painter’s creation or a generative algorithm’s output. Nevertheless, the scrutiny of fine arts presents demanding challenges due to the intrinsic variability and subjectivity inherent in the data. Notably, similar objects may be rendered in divergent artistic styles, while conversely, a singular artwork can evoke disparate emotional reactions among different observers. In response to this complexity, numerous models tailored for artwork analysis have embraced a multimodal approach, accommodating various data inputs such as images, text descriptions, or structured knowledge graphs. For example, in [13], we proposed a multimodal neural architecture adept at combining visual and contextual features from artworks, thereby facilitating the identification of their style and genre. Similarly, Bose et al. [14] presented a Transformer-based architecture to perform sentiment analysis within visual arts. This model effectively integrates visual features extracted from images with textual embeddings derived from painting descriptions.

However, it is noteworthy that a scarcity of scholarly literature addresses the automated detection of AI-generated artworks. In light of this, the present study contributes by introducing a novel dataset, which encompasses both authentic pieces of art, obtained through *ArtGraph* [13], and synthetic artistic images, sourced from *ArtiFact* [15]. Our study then assesses how well-established Deep Learning models can differentiate between them.

3. Materials

In order to conduct valuable experiments to recognize synthetic works of art, we designed a new large-scale dataset consisting of authentic and generated artworks. Specifically, we exploited two datasets: *ArtGraph* [13] and *ArtiFact* [15]. The former is a knowledge graph specializing in art, representing different concepts related to artworks and artists. It contains 116,475 art pieces, divided into 32 styles and 18 genres. Instead, *ArtiFact* is a large-scale dataset containing 2,496,738 images, comprising 946,989 authentic images and 1,531,749 fake images, concerning different domains, which include but are not limited to human faces, animals, vehicles, and art. To ensure significant diversity in the data, the authors have randomly sampled images from multiple data sources for authentic images, whereas, concerning synthetic content, they have exploited captions and image masks from the COCO dataset [16], which were passed to text2image and inpainting models. Furthermore, the authors employed various random seeds to sample normally distributed noise, which was then used to generate additional images using generative models. These images underwent cropping with a fixed ratio of 5/8, ensuring the crop size fell within 160 to 2048 pixels. Lastly, all the images were resized to a uniform resolution of 200×200 pixels and saved in JPEG format.

To collect data, we first extracted all the synthetic images of art within the *ArtiFact* dataset. This way, we gathered a set of images consisting of 37,775 synthetic works of art generated by multiple algorithms (e.g., Stable Diffusion [17] and StyleGAN [18]). Whereas, for authentic artworks, we sampled 50,000 artworks from *ArtGraph*, stratifying this process on the style attribute to avoid extreme class imbalance. Figure 1 shows two images representing each class within our dataset. Interestingly, there appears to be little distinction between the authentic and generated images. This observation underscores the primary motivation driving our study.

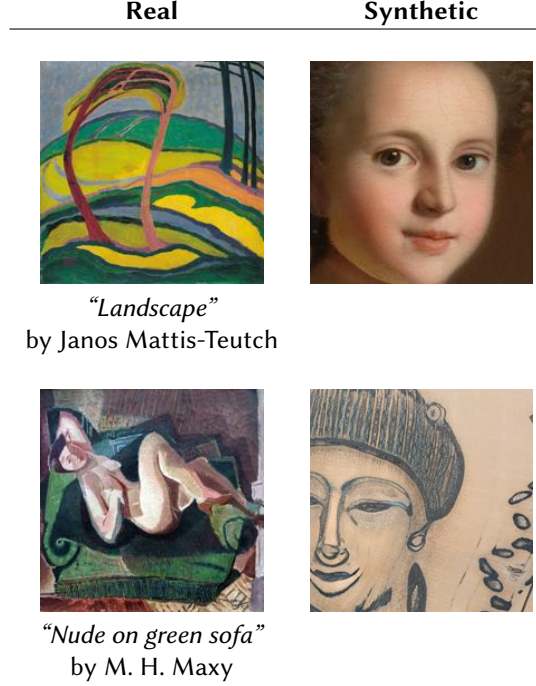


Figure 1: Example of artworks present in our dataset. On the left, authentic pieces of art sampled from *ArtGraph*. On the right, generated artworks available in *ArtiFact*. Distinguishing them only through visual inspection seems very challenging for a human.

4. Methods

This research presents a comprehensive Computer Vision framework centered around **binary classification algorithms**. To thoroughly evaluate the characteristics of this task, we conducted experiments using three cutting-edge techniques: VGG-19 [3], ResNet-50 [4], and ViT [5].

Figure 2 depicts the architecture of VGG-19, which we selected due to its capability to acquire hierarchical features from input images. This characteristic allows it to capture complex patterns and representations effectively. Moreover, the VGG-19 architecture is characterized by its simplicity and uniform structure, comprising convolutional and max-pooling layers. In Fig. 3, we present the conceptual design of the ResNet-50 model. This neural network introduces the concept of “residual learning”, **incorporating skip connections that facilitate the smoother flow of information throughout the network and address the vanishing gradient problem**. Lastly, Fig. 4 provides an overview of the ViT architecture, a state-of-the-art model recognized for its exceptional performance in various Computer Vision tasks, including image classification. ViT’s scalability makes it suitable for tasks involving images of varying sizes, as it allows for adjustments to the patch size and the number of layers, adapting to different input dimensions and computational resources.

In all cases, we leverage *transfer learning* to train the models mentioned above **by freezing all feature extraction layers and letting the model optimize just the last fully connected layers**, distinguishing whether the given image is an authentic artwork or a synthetic one. Accordingly,

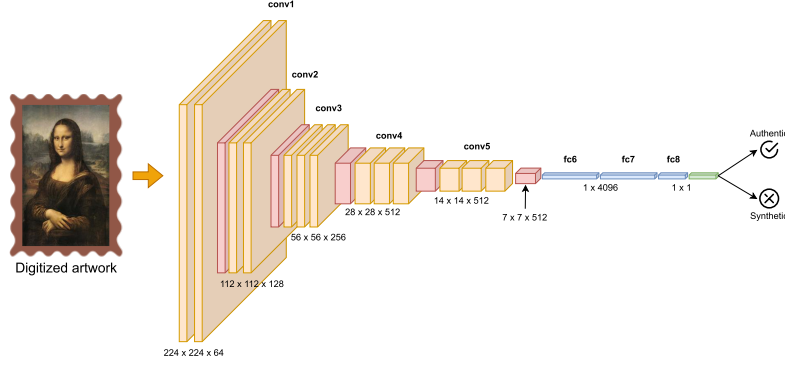


Figure 2: VGG-19.

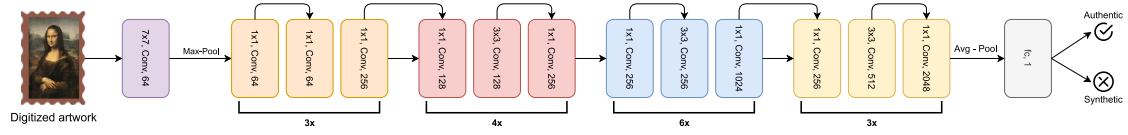


Figure 3: ResNet-50.

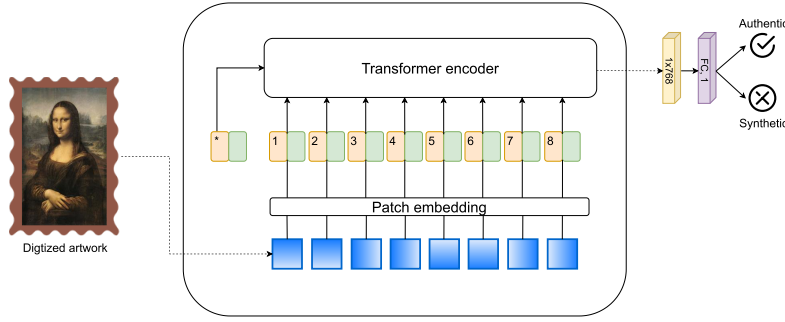


Figure 4: ViT.

once visual information is computed by each backbone, the feature vector is fed into a sigmoid-activated fully connected output layer, representing the probability that the given image is authentic. Then, to optimize the model, we use a binary cross-entropy loss function:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \ell(y_i, \hat{y}_i),$$

where N is the number of instances in the training set, y_i is the true label, \hat{y}_i is the model outcome, and $\ell(y_i, \hat{y}_i) = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i))$ is the classic binary cross-entropy loss.

We employed Grad-CAM, a technique introduced by Selvaraju et al. in [6], to enhance the explanatory aspect of our inference process. This approach allows us to shed light on the decision-making process of a typical Computer Vision method. Grad-CAM leverages the gradients associated with the network's output concerning the feature maps in the last

convolutional layer. As a result, we empower users to visualize the image patches that exert the most significant influence, which are highlighted in red. This feature amplifies the model’s usefulness, especially in a hypothetical collaborative system involving humans and AI, where informative automated suggestions can expedite decision-making.

5. Experiments

We performed some experiments for all the models mentioned above from both quantitative and qualitative points of view. In this section, we describe the experimental setup and discuss the obtained results.

5.1. Experimental Setup

The experiments were conducted using the Google Colab framework. Its resources include an Intel Xeon processor, 12 GB of RAM, and an NVIDIA T4 GPU with 15 GB of VRAM. All the models were implemented using the popular PyTorch library, which is well-known and suited for Computer Vision research.

We adopted an 80/10/10 splitting criterion, stratifying the partitioning on the target class. In this way, we ensured that the data distribution was preserved among training, validation, and test sets. It is worth noting that we used the same splitting to train all three models to have a fair comparison. Furthermore, all the images were preprocessed via center cropping. We resized them to a standard size of 224×224 pixels and performed normalization using the mean and standard deviation values of the ImageNet dataset. It is worth noting that all images (real and synthetic) were processed in the same way to conduct a fair experiment. Regarding the optimization stage, we used the Adam optimizer with a learning rate of 10^{-3} and a batch size of 32. An early stopping callback was introduced in the training loop with a patience of 3 and a learning rate schedule based on reducing it on the plateau to avoid overfitting.

To evaluate the effectiveness of the models, we used different performance metrics. Accuracy was calculated to show the percentage of well-classified instances. Furthermore, we calculated precision, recall, and F1-score to evaluate the model’s trustworthiness in identifying the authentic and the generated art pieces precisely.

5.2. Results

The quantitative results of the experimental analysis are shown in Table 1. The table shows that all three models are good predictors, with all metrics above 95%. The best model is ViT, but VGG-19 and ResNet-50 also perform well. These results highlight the ability of these models to detect subtle and imperceptible patterns that allow them to correctly distinguish between authentic and synthetic artworks, even though this would be a very challenging task for the naked eye.

Furthermore, in terms of qualitative evaluation, we assessed the models by analyzing their Grad-CAM outputs. Firstly, in Fig. 5, we present examples where the models correctly classified the images. Notably, VGG-19 emphasizes broad, coarse-grained features that span extensive areas within the given image. In contrast, ResNet-50 and ViT focus more on fine-grained details,

	Accuracy	Precision	Recall	F1
VGG-19	0.9581	0.9590	0.9562	0.9575
ResNet-50	0.9654	0.9645	0.9655	0.9650
ViT	0.9758	0.9752	0.9759	0.9755

Table 1
Experimental results.

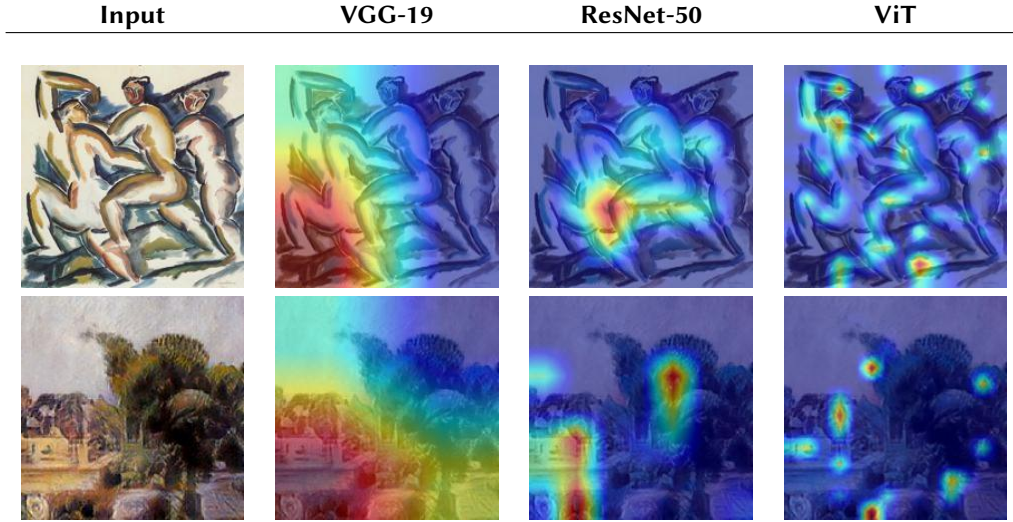


Figure 5: Examples of most discriminative regions using Grad-CAM for correctly classified images for real artwork (top) and synthetic art (bottom).

distinguishing whether a patch appears authentic or generated. In most cases, the models highlight image regions containing color discrepancies or lines that deviate from established patterns found elsewhere in the image. These discernible patterns are also noticeable in misclassified images, as depicted in Fig. 6. VGG-19, in these instances, emphasizes nearly half of the image, indicating confusion in its analysis. ResNet-50 directs attention to details in the sky where the pattern does not match the specific patch. Finally, ViT focuses on irregular lines within the eye. These observations highlight the inherent challenges in this task. Factors such as degradation over time or imperfections in the original artwork, such as inaccurately drawn lines by the artist, may contribute to these challenges.

From a conceptual standpoint, the suggested model accentuates image patches that deviate from the overall visual context. It considers broad and subtle features, including those that might not be immediately apparent to the naked eye. This discrepancy can be attributed to artifacts left in the image by generative models, which other algorithms can discern.

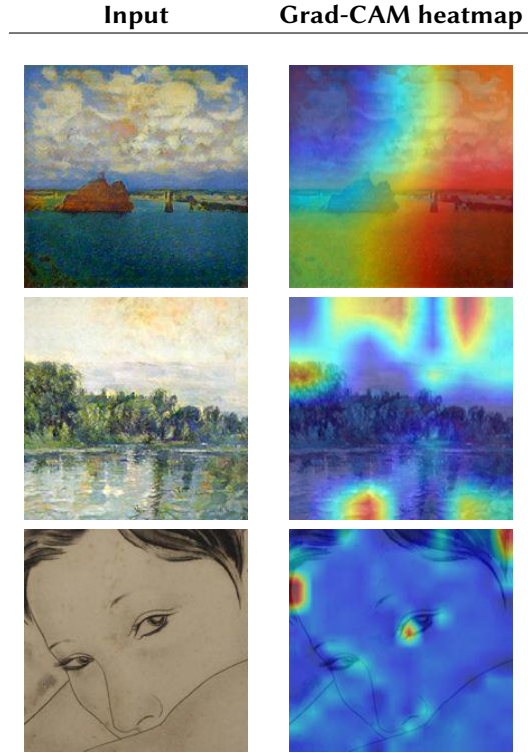


Figure 6: Examples of most discriminative regions using Grad-CAM for wrongly classified images using VGG-19 (first row), ResNet-50 (second row), and ViT (third row).

6. Conclusion

In this paper, our focus was recognizing artificial artworks generated by AI systems using Deep Learning models. To achieve this, we curated a novel dataset that included both authentic and generated artworks. We then experimented with various neural network architectures and conducted a comprehensive evaluation, considering both quantitative and qualitative aspects. Specifically, our quantitative analysis revealed that Vision Transformers exhibited strong predictive capabilities for this particular task, surpassing the performance of well-known models like VGG-19 and ResNet-50. Furthermore, it is worth noting that none of the three models misclassified the same test instances. However, upon closer examination of misclassified cases, we acknowledged that subtle imperfections at a fine-grained level could influence the models' ability to identify synthetic artworks correctly.

Future work could systematically compare the proposed method against other techniques designed for fake image detection, as the current work is exploratory in nature. Additionally, analyzing feature importance from a semantic perspective would improve model interpretability and move beyond pure visual interpretations of activation maps. Also, integrating contextual cues could help performance, as prior work has shown this modality to be effective for analyzing artistic metadata [13, 19]. Finally, ensemble approaches may be explored to harness the combined strengths of multiple models and enhance overall effectiveness for this task.

Acknowledgment

The research of Raffaele Scaringi is funded by a Ph.D. fellowship within the framework of the Italian “D.M. n. 352, April 9, 2022” - under the National Recovery and Resilience Plan, Mission 4, Component 2, Investment 3.3 - Ph.D. Project “Automatic analysis of artistic heritage via Artificial Intelligence”, co-supported by “Exprivia S.p.A.” (CUP H91I22000410007).

References

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, *Communications of the ACM* 63 (2020) 139–144.
- [2] F.-A. Croitoru, V. Hondru, R. T. Ionescu, M. Shah, Diffusion models in vision: A survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
- [3] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014).
- [4] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929* (2020).
- [6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [7] H. Li, B. Li, S. Tan, J. Huang, Identification of deep network generated images using disparities in color components, *Signal Processing* 174 (2020) 107616.
- [8] S. McCloskey, M. Albright, Detecting GAN-generated imagery using saturation cues, in: *2019 IEEE international conference on image processing (ICIP)*, IEEE, 2019, pp. 4584–4588.
- [9] X. Guo, X. Liu, Z. Ren, S. Grosz, I. Masi, X. Liu, Hierarchical fine-grained image forgery detection and localization, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3155–3165.
- [10] L. Guarnera, O. Giudice, S. Battiato, Level up the deepfake detection: a method to effectively discriminate images generated by GAN architectures and diffusion models, *arXiv preprint arXiv:2303.00608* (2023).
- [11] R. Amoroso, D. Morelli, M. Cornia, L. Baraldi, A. Del Bimbo, R. Cucchiara, Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images, *arXiv preprint arXiv:2304.00500* (2023).
- [12] Y. Hamid, S. Elyassami, Y. Gulzar, V. R. Balasaraswathi, T. Habuza, S. Wani, An improvised cnn model for fake image detection, *International Journal of Information Technology* 15 (2023) 5–15.
- [13] G. Castellano, V. Digeno, G. Sansaro, G. Vessio, Leveraging knowledge graphs and deep learning for automatic art analysis, *Knowledge-Based Systems* 248 (2022) 108859.

- [14] D. Bose, K. Somandepalli, S. Kundu, R. Lahiri, J. Gratch, S. Narayanan, Understanding of emotion perception from art, arXiv preprint arXiv:2110.06486 (2021).
- [15] M. A. Rahman, B. Paul, N. H. Sarker, Z. I. A. Hakim, S. A. Fattah, ArtiFact: A Large-Scale Dataset with Artificial and Factual Images for Generalizable and Robust Synthetic Image Detection, arXiv preprint arXiv:2302.11970 (2023).
- [16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.
- [17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-Resolution Image Synthesis With Latent Diffusion Models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.
- [18] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.
- [19] S. Aslan, G. Castellano, V. Digeno, G. Migailo, R. Scaringi, G. Vessio, Recognizing the emotions evoked by artworks through visual features and knowledge graph-embeddings, in: International Conference on Image Analysis and Processing, Springer, 2022, pp. 129–140.