

Project Report

GenRE: a supervised classification of fine art genres

Machine Learning for the Humanities – Academic year 2024-2025

Valentina Bertelli

1. Introduction

GenRe (Genre Recognition) investigates painting genre classification using a fine-tuned DenseNet121 model and compare its performance with a baseline based on feature extraction with DenseNet121 with pre-trained weights and k-nearest neighbour classification. Results were analysed using a confusion matrix and standard accuracy metrics for classification and Grad-CAM visualization for features interpretation.

2. State of the art

Advances in deep learning and computer vision, have opened new opportunities for the development of automatic tools that can support researchers in analysing and understanding visual arts. Large-scale digitization initiatives have led to an increasing availability of digitized visual art collections, which require appropriate metadata to be efficiently archived and retrieved. However, tagging large quantities of data is a time-consuming activity and, in the art domain, can be subjective and inconsistent due to different interpretations among experts. Machine learning facilitates a data-driven approach to art history and analysis and represent a faster and cheaper way of generating these types of metadata such as genre, style or authorship.

Genre recognition is a particularly relevant task in art history. The term "genre" refers to the traditional division of paintings based on the type of content they depict and is closely related to the semantic elements and scenes represented in artworks. Genre reflects artistic conventions through time and common elements shared by various artists. Accurate genre classification facilitates content-based search, digital curation, and comparative analysis of artworks.

3. Dataset

For this project the WikiArt dataset was chosen due to its large availability of images and metadata, as well as its extensive use in previous studies about machine learning-base art classification¹.

¹ See Castellano & Vessio (2021).

Given the large size of the original dataset, a smaller subset was obtained by filtering the [Hugging Face WikiArt](#) repository. Images were randomly selected regardless of style or author, using genre as the distinguishing characteristic, with fixed random seeds to ensure reproducibility.. The four chosen genres are: portrait, landscape, still life and sketch. The dataset was split into training, validation and testing sets, with a 70-15-15 ratio, resulting into 5600 images for training, 1200 for validation, 1200 for testing. All genres are equally represented across the three splits to avoid class imbalance.

Prior to training, images were pre-processed to ensure compatibility with the CNN architecture. All images were resized to a fixed resolution and converted to the RGB color space. Pixel values were normalized according to the ImageNet statistics. To improve generalization and reduce overfitting, data augmentation techniques such as random cropping and horizontal flipping were applied.

4. Methodology:

DenseNet121 was selected due to its strong performance and efficiency in fine art classification tasks². DenseNet uses dense skip connections that allow information to flow efficiently through the network, reducing the vanishing gradient problem.

The model was employed for feature extraction and classification and initialized with default weights obtained from pre-training on ImageNet. Since paintings genres are defined by subjects and scenes depicted, colour and texture features that are captured by the lower level of the model, don't contribute to genre classification as much as they do in style or artist classification tasks³. This justified the choice of a partial fine-tuning of the model: the lower layers were frozen, while the final dense block and the last normalization layer were fine-tuned on the training dataset. Cross-entropy loss was used as the objective function to optimize its classification performance, and the Adam optimizer was used for parameter optimization⁴.

The model was trained for a maximum of 20 epochs. Given the high computational power required for fine-tuning higher-level feature extraction layers, early stopping was applied to prevent overfitting and unnecessary computation. Training was stopped when the validation loss did not improve for five consecutive epochs.

² See Boyadzhiev et al. (2025). Densenet121 resulted as the most accurate and efficient architecture, compared to other models.

³ See Cetinic & Grgic (2016). The study also confirmed that combining different types of features is useful style or artist classification, but doesn't significantly improve genre classification.

⁴ Mezina & Burget (2025) confirmed that these loss function and optimizer, as appropriate for art classification tasks.

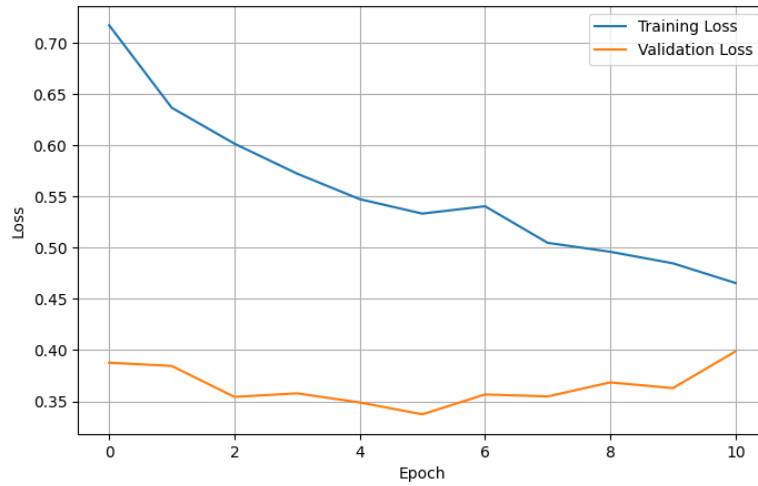


Figure 1 The graph shows the behaviour of training and validation

Training and validation curves show a stable convergence behaviour, with validation accuracy reaching approximately 88%. The gap between training and validation performance remains limited, indicating good generalization and no significant overfitting. Due to stochastic training and data augmentation, the exact epoch at which early stopping occurs may vary slightly between runs. From this point on the fin-tuned model will be referred as GenRe.

To assess the effectiveness of the fine-tuning strategy, a baseline model was implemented for comparison⁵. The baseline consists of a DenseNet121 network with fixed ImageNet pre-trained weights and a k-nearest neighbors (kNN) classifier. Unlike the GenRe, the baseline does not update feature representations during training. It was used to extract features from the same three datasets used for the previous model. Since kNN classifies images based on their similarity in the feature space, it is a well suited classifier for evaluating the quality of fixed deep features, without introducing additional learning in the model. Although the feature extractor is not trained, kNN still requires the selection of the hyperparameter k. For this reason, features from the training set were used to fit the classifier, while the validation set was used to select the best k value based on accuracy, ensuring a fair and reliable comparison. The baseline relies exclusively on generic visual representations, allowing a direct comparison with GenRe, where feature representations are adapted to the painting genre domain.

5. Results and conclusions:

The test dataset was used to evaluate both the GenRe and the baseline. Classification results, where compared using confusion matrices and standard evaluation metrics. Than Grad-CAM visualization was applied to access how fine-tuning affected features extraction.

⁵ The same strategy has been implemented by Boyadzhiev et. Al. (2025).

Both models achieved good classification performances, confirming DenseNet121 effectiveness when applied to the art domain. GenRe showed slightly superior results, confirming the usefulness of fine-tuning and suggesting that a deeper training of the lower layers of the model would lead to even better results. In both cases, the sketch class was the most difficult to identify (248 correct predictions for GenRe, 239 for the baseline), and it was often confused with the portrait class (26 times by GenRe, 31 by the baseline). This behaviour was expected due to the fact that most sketch images represents people, making the roughness of the drawing the main discriminative feature.



Figure 2 Confusion Matrix from fine-tuned model

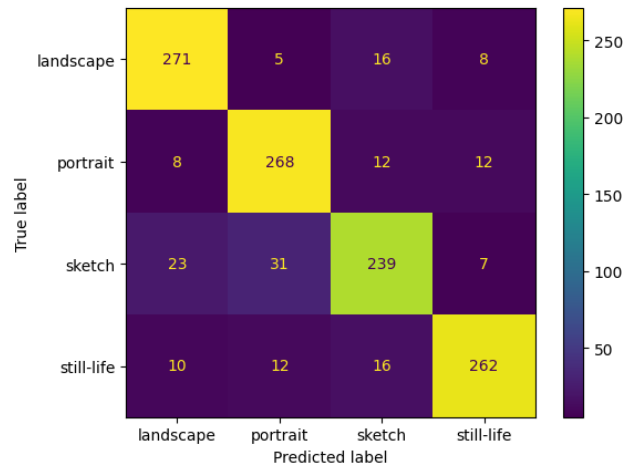


Figure 3 Confusion Matrix from baseline model

Standard classification metrics (Accuracy, Precision, Recall and F1-score) were used to evaluate performance. Overall, the results confirm GenRe’s superiority.

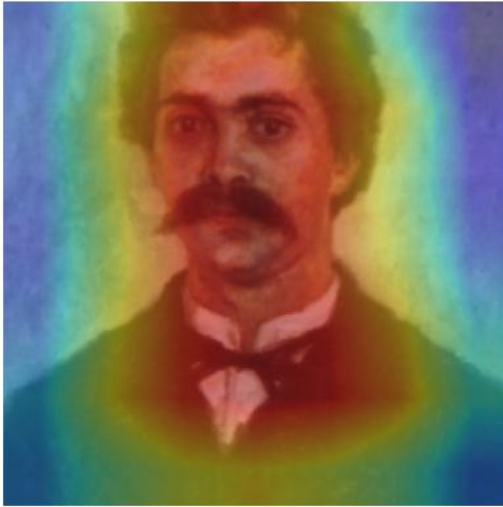
| Model | Accuracy | Precision | Recall | F1-score |
|----------|----------|-----------|--------|----------|
| GenRe | 0.8933 | 0.8931 | 0.8933 | 0.8930 |
| Baseline | 0.8666 | 0.8669 | 0.8666 | 0.8663 |

Table 1 Results of evaluation metrics of GenRe and the baseline model

Grad-CAM was used to interpret models’ decisions by visualizing the regions of the image that contributed most to the predictions. Most images were classified correctly by both models, but with some significant differences in the identification of relevant features.

The baseline model showed a tendency to focus on small, localized regions of the image and object fragments, having more difficulties in understanding the object’s semantic role in the composition, for example misclassifying portraits when the background was too detailed or when the subject was depicted in a full-body composition. GenRe appears to have learned more effectively what elements distinguish a genre from the other. In portrait paintings correctly classified, the baseline often identified discriminative features in the area underneath the subject’s neck, whereas GenRe consistently focused on the subject’s face (e.g. Figures 4 and 5).

GradCAM of fine-tuned model's features
Predicted genre: portrait



GradCAM of baseline model's features
Predicted genre: portrait

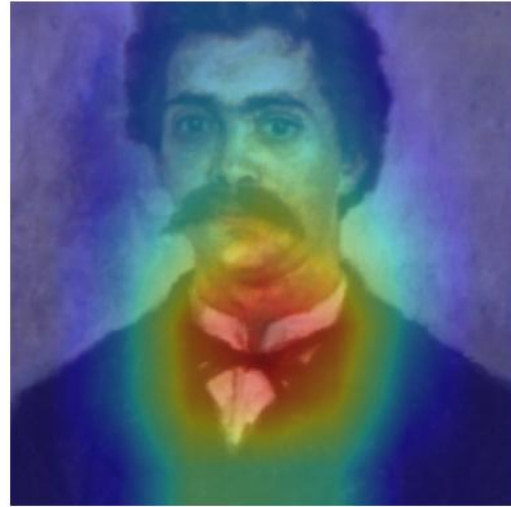


Figure 4 Grad-CAM of a portrait from the fine-tuned model Figure 5 Grad-CAM of a portrait from the baseline model

In the majority of landscape painting, the baseline model tended to focus on foreground objects, when present, or isolated background areas, while the fine-tuned model's activations were more widely distributed across the background, highlighting natural elements such as mountains or woods, or concentrating on the perspective focal point. Looking at Figure 6 and 7, for example, both models classified the image correctly, but the fine-tuned model focused on natural elements such as the cliff on the left side of the background and the clouds in the upper-right corner, while the baseline mainly concentrated to the ship in the foreground.

GradCAM of baseline model's features
Predicted genre: landscape



GradCAM of fine-tuned model's features
Predicted genre: landscape



Figure 6 Grad-CAM of a landscape from the fine-tuned model Figure 7 Grad-CAM of a landscape from the baseline model

Misclassification in the sketch genre class were the most common for both models, as shown in the confusion matrices. Grad-CAM visualizations suggest that sketches were correctly classified when discriminative feature identified when small discriminative regions characterized by dark shadows and simple line compositions, such as sketched contours or shading around human figures (e.g. Figure

8). When instead the model's attention shifted toward depicted subject rather than these structural features, sketches images were assigned to more figurative genre (e.g. Figure 9). This suggests that since sketches' most discriminative cues are more related to low-level features, deeper fine-tuning could be fundamental to further improve performance for this genre.

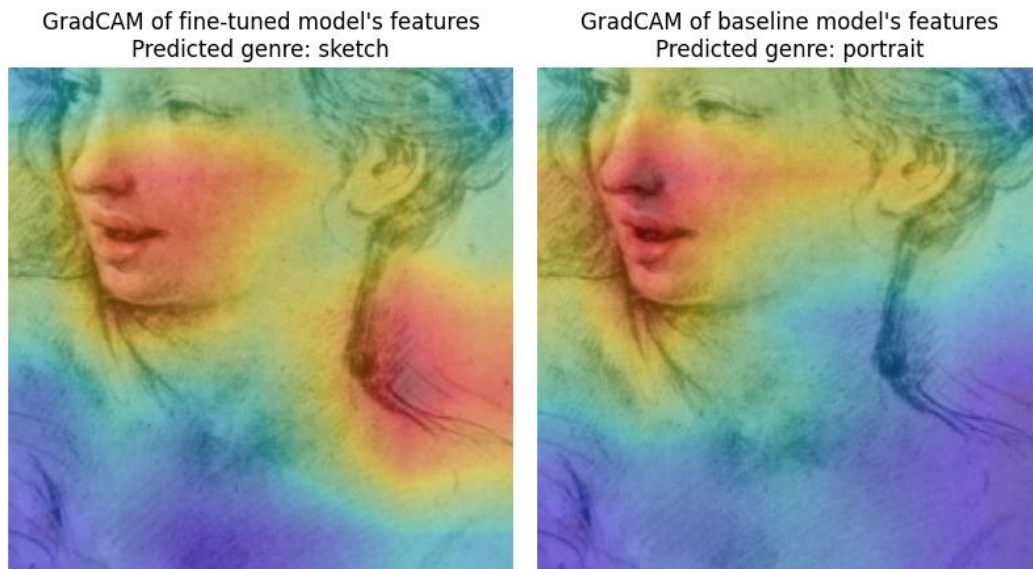


Figure 8 Grad-CAM of a sketch from the fine-tuned model Figure 9 Grad-CAM of a sketch from the fine-tuned model

As show in the confusion matrices, still-life images were the class in which GenRe achieved the largest improvement compared to the baseline. In some cases, the baseline produced clearly incorrect predictions, such as classifying object-centric paintings as portraits or landscapes, with no evident visual explanation (e.g. Figure 11). In contrast, GenRe demonstrated a more consistent ability to distinguish object-based paintings, indicating a better understanding of still-life specific visual cues (e.g. Figure 10).

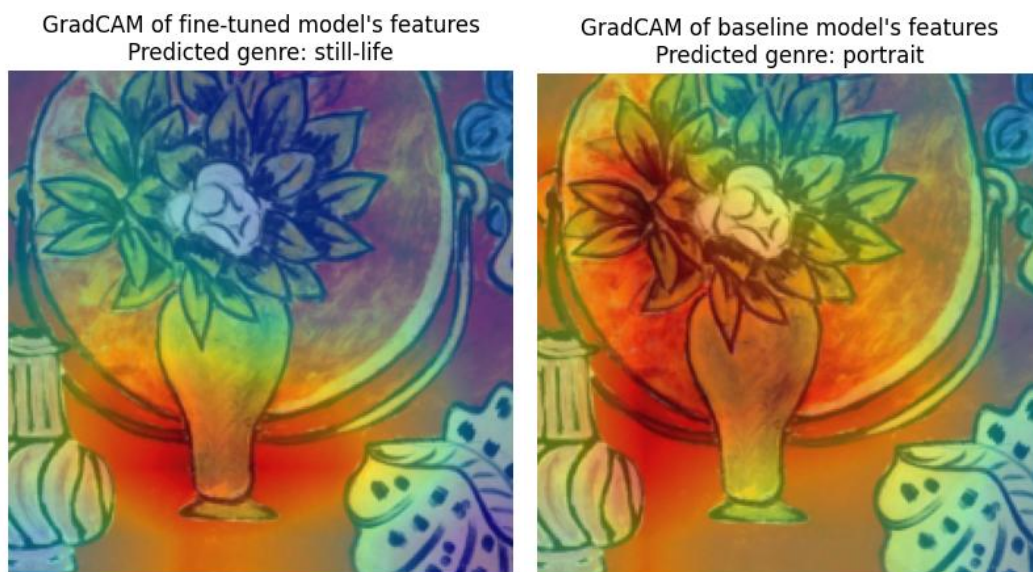


Figure 10 Grad-CAM of a still-life from the fine-tuned model Figure 11 Grad-CAM of a still-life from the fine-tuned model

6. Bibliography:

- Boyadzhiev T., Lagani G., Ciampi L., Amato G., Ivanova K. (2025). 'Comparison of Different Deep Neural Network Models in the Cultural Heritage Domain', arXiv:2504.21387, preprint, arXiv <https://doi.org/10.48550/arXiv.2504.21387>.
- Castellano, G., Vessio, G. (2021). Deep learning approaches to pattern extraction and recognition in paintings and drawings: an overview. *Neural Comput & Applic* **33**, 12263–12282. <https://doi.org/10.1007/s00521-021-05893-z>.
- Cetinic E. & Grgic S. (2016). 'Genre Classification of Paintings', *International Symposium ELMAR*, IEEE, September 2016, 201–4, <https://doi.org/10.1109/ELMAR.2016.7731786>.
- Li W. (2025). 'Enhanced Automated Art Curation Using Supervised Modified CNN for Art Style Classification', *Scientific Reports* 15, no. 1 (2025): 7319, <https://doi.org/10.1038/s41598-025-91671-z>.
- Mezina A. & Burget R. (2025). 'EnsArtNet: Ensemble Neural Network Architecture for Identifying Art Styles from Paintings', *Journal of Cultural Heritage* 72 (March 2025): 71–80, <https://doi.org/10.1016/j.culher>.
- Song H., Wang L., Song C. (2025). 'Research on the Analysis of Image Characteristics in Fine Art Painting under the Application of Machine Learning Technology', *Ain Shams Engineering Journal* 16, no. 8: 103456, <https://doi.org/10.1016/j.asej.2025.103456>.
- Zhao W, Zhou D, Qiu X, Jiang W (2021) Compare the performance of the models in art classification. *PLoS ONE* 16(3): e0248414. <https://doi.org/10.1371/journal.pone.0248414>.