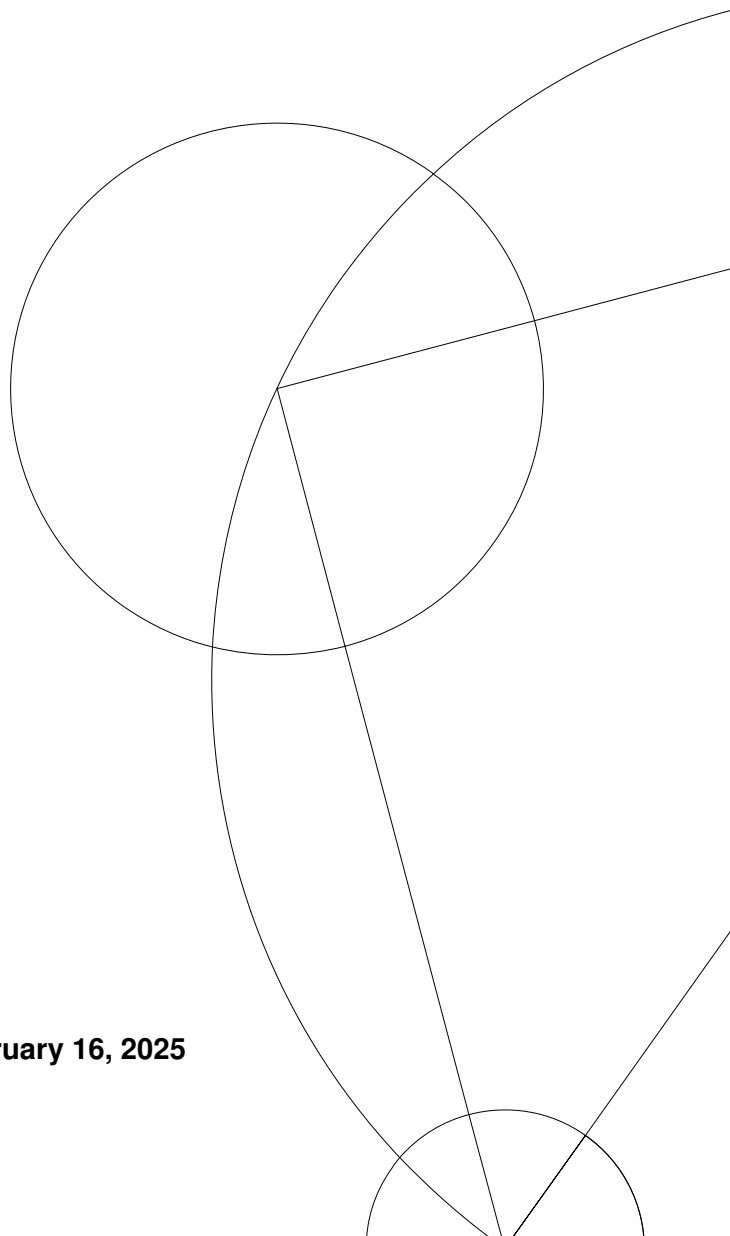
Computational Cognitive Science II

Andrea Schröter (kbn999)

Multimodal Painter Attribution:

Integrating Visual and Textual Features for Improved Classification Performance



Tutor: Patrizia Paggio, Costanza Navaretta

Department of Nordic Studies and Languages: February 16, 2025

Abstract

Painter classification describes the process of automatically attributing an artwork to an artist, and is challenging due to the need for additional contextual knowledge and art historical expertise. This study examines the effectiveness of multimodal approaches to painter classification by comparing the performance of models using joint textual-visual features with those using only unimodal features. For the unimodal baselines, visual features are extracted using a ResNet100 model on paintings from the SemArt data set, whereas a pre-trained BERT is trained on the respective artistic comments. Joint features are extracted with VilBERT, integrating textual and visual features. In line with the assumption that including additional modalities increases classification accuracy, the model with joint feature representations outperforms the unimodal baselines and even exceeds state-of-the-art accuracies in multimodal painter classification. Thus, the integration of textual and visual features can facilitate and enhance automated art analysis, particularly in the domains of painter attribution and identification.

1 Introduction

Painter classification describes the task of automatically attributing an artwork to a painter using machine learning techniques. The increasing digitisation of paintings and the development of art-historic archives have led to an increased demand for methods to process and analyze artwork, including tasks such as image retrieval, genre, style, and painter classification.

However, painter attribution is considered considerably more challenging than the classification of natural images. One reason for this is the varying degree of abstractness and symbolism, making expertise and a background in art history necessary to correctly analyze and categorize artwork [1, 14]. Furthermore, artists may change their painting style, and on the other hand, several painters can follow similar painting styles [9].

While image classification has been extensively studied in the recent years, not many studies focus on automatic art analysis. The majority of studies conducted in this field solely extract visual features from the paintings to classify its painter, genre, and style [4, 10, 9, 1]. Typically, Convolutional Neural Networks (CNNs) are used. For instance, Tan et al. [14] fine-tune a CNN pre-trained on the ImageNet dataset for painter classification, and David et al. [4] combine a convolutional autoencoder and a CNN to extract and classify visual features.

However, visual features may not be sufficient for the successful interpretation of art. In their study, Garcia et al. [6] emphasise the significance of additional contextual information regarding art history, artistic styles and depicted subjects. Following this approach, Zhao et al. [17] propose utilising artistic comments on paintings for classification. The authors compare computer vision and NLP techniques for painter classification and show that applying NLP techniques to artistic comments achieves similar or even higher accuracies[1] compared to ResNet models (a type of CNN) using pixels of the paintings. Nevertheless, multimodal approaches to painter classification or art analysis are rare and mainly limited to using knowledge graphs to leverage metadata and contextual information from artistic comments [7, 3]. For instance, Castellano et al. [3] integrate visual features and embeddings from a knowlegde graph based on WikiArt, which contains information about artworks, artists, style, genre, and relationships to other painters and project them into the same multidimensional space. Their study highlights the potential of multimodal approaches in art classification.

However, the integration of embeddings generated from a knowledge graph can result in an increased computational cost. Since artistic comments alone have proven to result in high accuracies ([17]), the paper at hand focuses on integrating contextual features obtained from only artistic comments with visual features to train a multimodal model for painter classification. The objective of this study is hence to determine whether models that integrate visual (the painting) and textual features (artistic comments) demonstrate higher classification performance in artist prediction compared to models relying on textual or visual features only.

Based on the tendency of multimodal models to outperform unimodal ones [13] and the success of models leveraging artistic comments and contextual information [16, 7, 3], it is hypothesized that the joint embeddings will outperform the baselines. To test this, visual features will be extracted by applying a pre-trained state-of-the-art CNN, a ResNet100 model. These visual features will be passed as an input to VilBERT, a state-of-the-art vision-and-language model, to obtain joint visual-textual representations. Textual features will be obtained using BERT. All features will be trained on separate neural classifiers with a simple Multi Layer Perceptron (MLP) architecture, and a Wilcoxon signed-rank test will be conducted on the test accuracies to evaluate whether the multimodal model significantly outperforms the baselines.

2 Methodology

2.1 Data

For this study, the SemArt dataset for semantic art understanding was utilized [8]. This data set consists of 21,383 images of paintings, accompanied by a comment or description and various attributes associated to the the paintings (Author, Title, Date, Technique, Type, School and

[1] High accuracies of 0.702 were obtained by ArtGCN, an art graph convolutional network trained on an artistic comment graph containing co-occurence and other types of document relations [17].

Timeframe). In total, it comprises paintings from 3,281 different painters spanning the period between 801 and 1900.[2]

To ensure a balanced data set regarding the number of paintings per artist and due to limitations in computational resources available, only the ten most frequent artists were included. Subsequently, undersampling with regard to the least frequent of the painters was performed. The balanced dataset thus contained 178 paintings each by the ten painters Van Gogh, Rembrandt, Giotto, Tiziano, El Greco, Cranach, Veronese, an unknown Italian master, and Tiepolo.[3]

As some comments mentioned the painter of the respective painting, while others did not (see two examples of the data set in Appendix B), Named Entity Recognition (NER) was applied to mask the painters' names. This was done as an attempt to increase the homogeneity of the comments in the dataset and to prevent explicit mentions of the painters to be a cue. Following the original train-val-test split, a 90-5-5 ratio of training, validation, and testing set was used, resulting in 1602 training, 89 validation, and 89 test data points.

2.2 Feature extraction and Models

To extract the visual embeddings, Detectron2's [15] Mask R-CNN ResNet101-FPN model[4] was applied to detect regions in the paintings containing objects. These regions were then further processed to obtain features capturing visual characteristics of the objects.

The visual embeddings were then fed to VilBERT [12] to create joint visual-textual feature representations. VilBERT (Vision-and-Language BERT) is an extension of the BERT architecture that allows for the learning of joint representations of natural language and images. The joint representations of paintings and their corresponding comments were obtained by extracting the last hidden layer of the VilBERT model.

As the BERT architecture is integrated into VilBERT, to allow for better comparability the comments were then encoded using a BERT tokenizer and then passed to a simple neural network architecture containing the bert-base-uncased model pre-trained on Wikipedia and an English book corpus [5]. The models' weights were frozen to prevent them from training, and an intermediate dense layer with 512 neurons and ReLU activation function, together with an output classification layer with 10 neurons and the Softmax activation function were stacked on top.[5]

The visual and joint features were fed into two MLPs, consisting of two dense layers of 128 and 64 neurons, respectively, with ReLu as the activation function. After each dense layer, a batch normalisation and a dropout regularisation of 0.5 were applied to prevent overfitting. A classification dense layer with Softmax activation function was stacked on top. All three classifiers were trained with ten different random seeds for 15 epochs at a learning rate of 10^{-4}. Adam was used as the optimiser and CategoricalCrossentropy as the loss function.[6] Test and validation accuracies were reported for each model.

2.3 Evaluation and Statistical Analysis

To evaluate the classifiers' performances, accuracy was used as a metric. Additionally, a one-sided Wilcoxon signed-rank test was conducted to analyse whether the differences in test accuracy distributions obtained by the baseline models and the multimodal models are significant.

[2]The data was collected from the Web Gallery of Art (WGA) `https://www.wga.hu/`, an online art catalogue providing contextual comments about art paintings.

[3]The datasets and the code used for data preprocessing, feature extraction, and for training the models can be found in the following Github repository: `https://github.com/reascr/Multimodal_Painter_Attribution`.

[4]The model was pre-trained on the COCO data set for object recognition [11].

[5]Making the architecture more similar to the one used for the visual and joint embeddings by adding intermediate layers resulted in very poor accuracies, suggesting that the architecture got too complex for the multiclass classification problem at hand. For this reason, the architecture used to trained the textual embeddings was kept simpler.

[6]Note that for the visual and joint embeddings, Sparse Categorical Crossentropy was used as the loss function because the labels were not onehot encoded.

3 Results

Table 1 shows the accuracies obtained by evaluating the three models on the test set. The model trained on visual features exhibited the highest accuracy of 0.472, which is comparable to the highest accuracy observed for the BERT model (0.471). The model trained on joint visual-textual representations outperformed both baselines, with accuracies ranging between 0.573 and 0.708. The validation accuracies during training the models on the random seed 4[7] are visualized in Figure 1. The model trained on joint features reaches an accuracy of 0.7 after epoch 9, while the unimodal models perform less well, with slightly above 0.3 for the visual features and slightly above 0.4 for the model trained on textual features).

Table 1: Test set accuracies for all random seeds and models.

Seed	Visual Features	Text Features	Joint Features
0	0.438	0.438	0.663
1	**0.472**	0.438	0.607
2	0.449	0.382	0.640
3	0.416	0.416	0.640
4	0.360	0.416	0.652
5	0.393	0.438	0.652
6	0.427	0.438	0.674
7	0.404	0.438	0.685
8	0.461	**0.471**	0.573
9	0.449	0.382	**0.708**

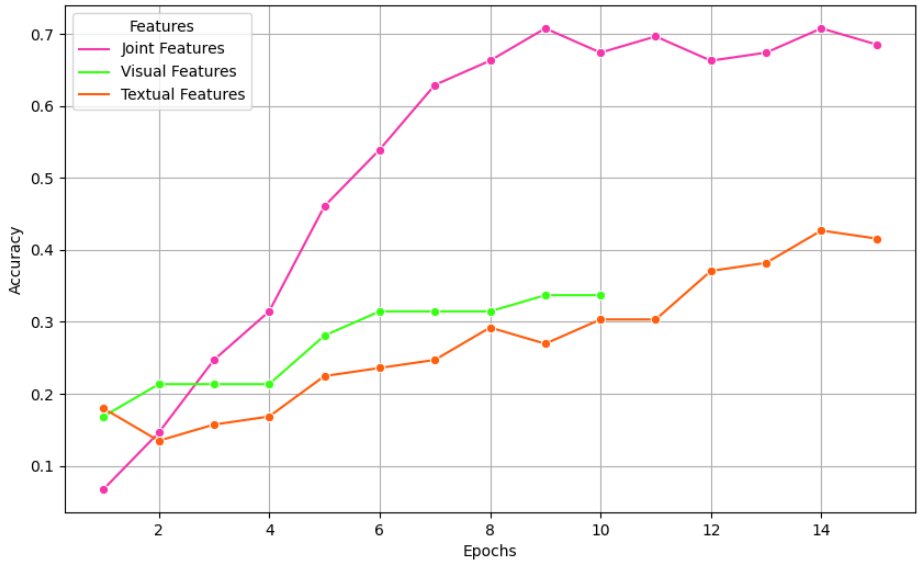


Figure 1: Validation accuracies per epoch using a random seed of 9. Note that early stopping was implemented in the MLPs for the visual and joint features, resulting the model trained on visual features to end the training early.

The Wilcoxon signed-rank test yielded a p-value of 0.00098 for both baselines, indicating that the model using joint features significantly outperformed both the textual as well as the visual baseline. Given that the p-value was identical for applying the test onto the multimodal model and each of the baselines, an additional Wilcoxon signed-rank test was conducted to ascertain whether the test accuracies obtained by training on visual and textual features were

[7]Validation accuracies for all random seeds and epochs can be found in the GitHub repository: `https://github.com/reascr/Multimodal_Painter_Attribution/tree/main/wilcoxon_signed_rank_test`.

likely to be drawn from the same distribution. The resulting p-value was 1.0, indicating that the distributions of test accuracies were similar for both baselines.

4 Discussion

The results demonstrate that joint visual-textual features significantly enhance the performance in painter classification, outperforming the unimodal baselines. Notably, the simple MLP trained on joint features obtained by VilBERT reaches a higher accuracy (0.708) than the state-of-the-art results using knowledge graphs (0.622 and 0.625, as shown in Table 2).

Table 2: Comparison of methods, training data, and accuracy of previous studies.

Features	Method (for Feature Extraction)	Dataset	n labels	n artworks	Accuracy
Multimodal	ContextNet Knowledge Graph [7]	SemArt	3166	19,244	0.622
	ArtGraph Knowledge Graph [3]	WikiArt	300	63,145	0.625
	Proposed (ViLBERT)	SemArt	10	1602	**0.708**
Textual	RoBERTa [17]	SemArt	3166	19,244	0.465
	ArtGCN [17]	SemArt	3166	19,244	**0.702**
	Proposed (BERT)	SemArt	10	1602	0.471
Visual	CNN-finetuned [14]	WikiArt	23	20,000	**0.761**
	ResNet101 [17]	SemArt	3166	19,244	0.519
	Proposed (ResNet101)	SemArt	10	1602	0.472

However, the results obtained for the baselines are significantly lower than those of state-of-the-art approaches (e.g. the ArtGCN proposed by Zhao et al. [17]). One possible explanation for the low accuracy of visual features is that the ResNet101 model used to extract them was pre-trained on the COCO data set. As Banerji and Sinha [2] state, pre-training CNNs on different image datasets might not be suitable for painter classification. Moreover, the unbalanced distribution of types and schools in the data set could have further complicated the classification problem. For instance, the Italian school and the religious type were almost twice as common as other categories (see Figures A3 and A4). Examining the confusion matrix for the model trained on visual features, however, a clear tendency to confound Italian painters with each other could not be detected[8]. Upon examination at the validation accuracy and loss, clear signs of overfitting can be detected for the visual features (Figure A6) as the validation accuracy is decreasing after the 5th epoch, while the training accuracy continues to increase. This trend cannot be observed for the textual and joint features in this extend (see Figure A5 and A7). The comparatively low performance of BERT for painter classification may be attributed to the limited training data size, the heterogeneity of the comments, which varied in detail, content, and length, and the inappropriate hyperparameter selection.[9]

However, a direct comparison with previous studies is difficult due to differences in datasets and sample and label sizes. In addition, the model using joint features was not trained on the whole undersampled data set (1500/1602 paintings) due to limitations in computing power. Further limitations concern the limited data size, the imbalance of the data set regarding school and type, and the limited experiments conducted to improve the accuracy of the models by fine-tuning their hyperparameters and changing the models' architectures. Future studies should address these shortcomings to further enhance the classification accuracy for joint features. Nevertheless, the objective of this study was to compare the performance of models using multimodal features with those using unimodal features, rather than achieving state-of-the-art accuracies for the classification problem at hand.

[8] Although Giotto, an Italian painter, is confounded with the unknown Italian master, the Spanish painter El Greco is equally confounded with the Dutch painter Rubens. Note that it would have been preferable to exclude the unknown Italian master from the painter classification task at hand, since it was unclear whether his paintings seemed to stem from one anonymous painter or whether the origin of the paintings are not clear in general [8].

[9] Moreover, masking of the painter's name by NER recognition might not have detected all painter's names.

One potential application of the suggested multimodal approach to painter classification is the task of painting authentication, which involves determining whether a given a given painting was painted by a specific artist. Furthermore, the combination of artistic comments with additional contextual information could be employed to identify the painter of an artwork.

5 Conclusion

This paper investigated the effectiveness of using joint visual-textual feature representations for training models for painter classification on the SemArt data set. For the unimodal baselines, visual features were extracted from images of paintings using a ResNet101 model, while textual features were obtained using BERT. The integration of visual and textual features into multimodal features was achieved through the use of VilBERT. All three feature representations were fed into simple MLPs and test accuracies were reported. In line with the assumption that adding a modality dimension enhances classification accuracy, further emphasized by the promising results of multimodal approaches in art classification, the model trained on joint features significantly outperformed the unimodal baselines. These results suggest that artistic comments provide supplementary information that aid in predicting the painter of an artwork.

One shortcoming of this paper concerns the relatively small size of the data set. In addition, the artistic comments were heterogeneous in terms of content, length, and detail and the subsample drawn from the SemArt data set was biased towards Italian painters and religious paintings. Future studies should address these limitations by increasing the sample size, using different data sets, and selecting appropriate hyperparameters and model configurations. Despite the limitations, this study highlights the potential of multimodal approaches in painter classification and automatic art analysis.

References

[1] S. Agarwal, H. Karnick, N. Pant, and U. Patel. Genre and style based painting classification. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 588–594. IEEE, 2015.

[2] S. Banerji and A. Sinha. Painting classification using a pre-trained convolutional neural network. In *Computer Vision, Graphics, and Image Processing: ICVGIP 2016 Satellite Workshops, WCVA, DAR, and MedImage, Guwahati, India, December 19, 2016 Revised Selected Papers*, pages 168–179. Springer, 2017.

[3] G. Castellano, G. Sansaro, and G. Vessio. Integrating contextual knowledge to visual features for fine art classification. *arXiv preprint arXiv:2105.15028*, 2021.

[4] O. E. David and N. S. Netanyahu. Deeppainter: Painter classification using deep convolutional autoencoders. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*, pages 20–28. Springer, 2016.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] N. Garcia, B. Renoust, and Y. Nakashima. Understanding art through multi-modal retrieval in paintings. *arXiv preprint arXiv:1904.10615*, 2019.

[7] N. Garcia, B. Renoust, and Y. Nakashima. Contextnet: representation and exploration for painting classification and retrieval in context. *International Journal of Multimedia Information Retrieval*, 9(1):17–30, 2020.

[8] N. Garcia and G. Vogiatzis. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[9] K.-L. Hua, T.-T. Ho, K.-A. Jangtjik, Y.-J. Chen, and M.-C. Yeh. Artist-based painting classification using markov random fields with convolution neural network. *Multimedia Tools and Applications*, 79:12635–12658, 2020.

[10] M. O. Kelek, N. Calik, and T. Yildirim. Painter classification over the novel art painting data set via the latest deep neural networks. *Procedia Computer Science*, 154:369–376, 2019.

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and 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*, pages 740–755. Springer, 2014.

[12] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

[13] Z. Lu. A theory of multimodal learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] W. R. Tan, C. S. Chan, H. E. Aguirre, and K. Tanaka. Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. In *2016 IEEE international conference on image processing (ICIP)*, pages 3703–3707. IEEE, 2016.

[15] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[16] W. Zhao, D. Zhou, X. Qiu, and W. Jiang. Compare the performance of the models in art classification. *Plos one*, 16(3):e0248414, 2021.

[17] W. Zhao, D. Zhou, X. Qiu, and W. Jiang. How to represent paintings: A painting classification using artistic comments. *Sensors*, 21(6), 2021.

Appendix

A Examples of Paintings and Meta Data in the SemArt Data Set

Title: Landscape with a Shepherd and a Shepherdess
Author: TENIERS, David the Younger
Type: landscape
School: Flemish
Timeframe: 1651-1700

This pastoral genre piece depicts a shepherd and a shepherdess and their flock along a path in an extensive Arcadian landscape.

Figure A1: Example of a landscape painting and corresponding metadata (title, author, type, school, and timeframe) in the SemArt data set.

Title: Flora
Author: TIZIANO Vecellio
Type: portrait
School: Italian
Timeframe: 1501-1550

This is one of Titian's most beautiful works, which, in the warm and impassioned intensity of the colour, sums up the youthful period of Titian. The beautiful woman carrying flowers is thought to be Flora, the classical goddess of flowers and spring. The title of Flora goes back to an engraving which was made from the picture in the 17th century by Sandrart. This painting is one of the first of a series of portraits of ideal female beauty that Titian painted. The sheen of her reddish golden hair, the soft hue of her skin, and the just visible breast whose bareness is skillfully emphasized by her hand and the pink brocade, display Titian's abilities as a subtle colourist and his sure feeling for sensuality.In the 17th century, the Flora came to the Netherlands and inspired Rembrandt to paint his wife Saskia in the same guise, albeit less scantily clad.

Figure A2: Example of a portrait painting and corresponding metadata (title, author, type, school, and timeframe) in the SemArt data set.

B Distribution of Meta Data in the Balanced Data Set

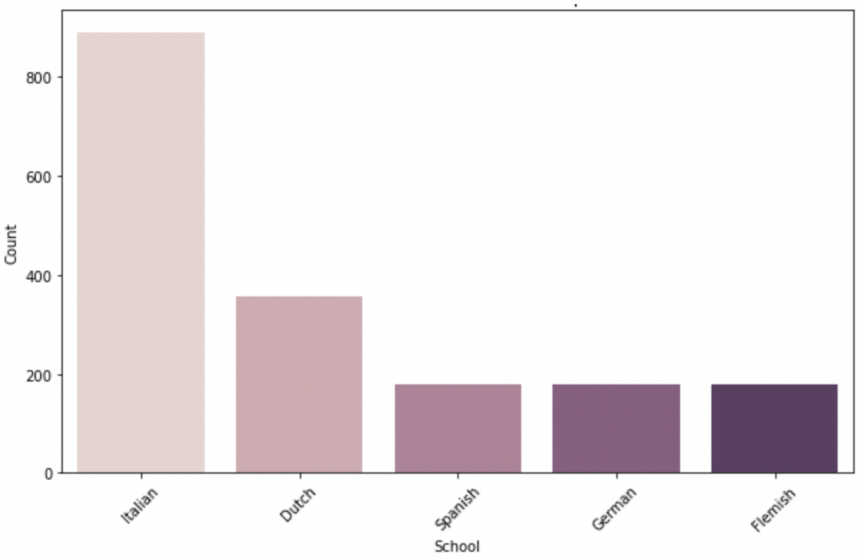


Figure A3: Distribution of schools among painters in the balanced and undersampled data set.

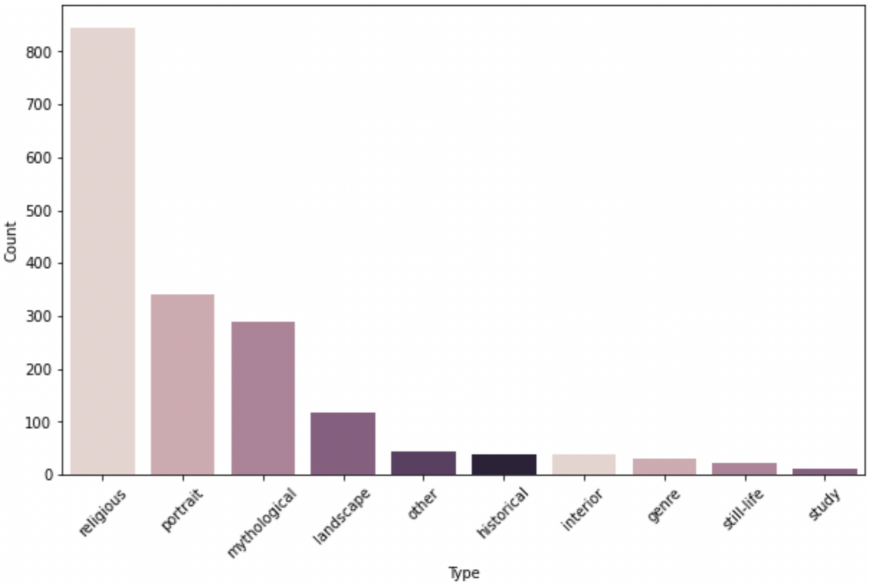


Figure A4: Distribution of painting types in the balanced and undersampled data set.

C Validation Accuracies during Training

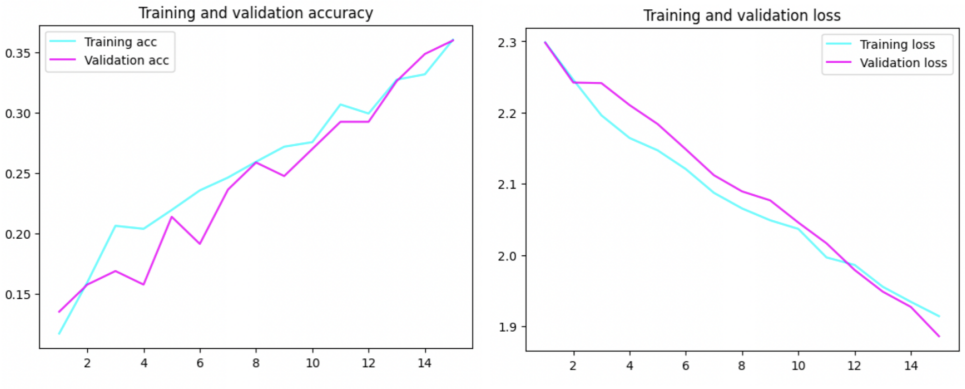


Figure A5: Training and validation accuracy and loss for the model trained with textual features, utilizing a random seed of 9.

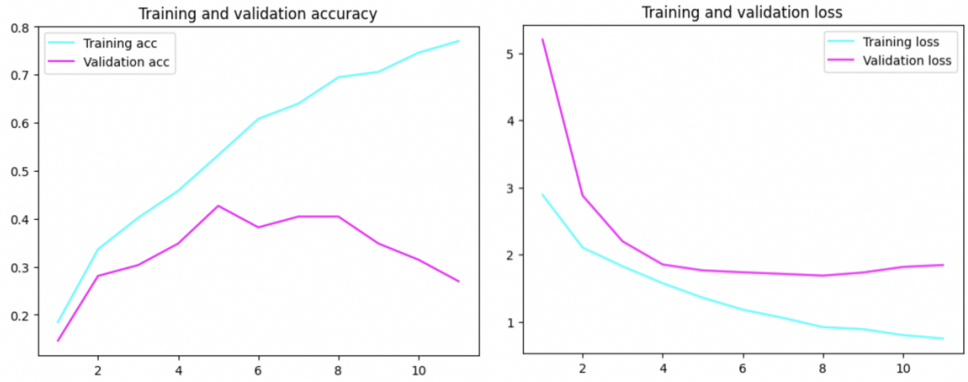


Figure A6: Training and validation accuracy and loss for the model trained with visual features, utilizing a random seed of 9.

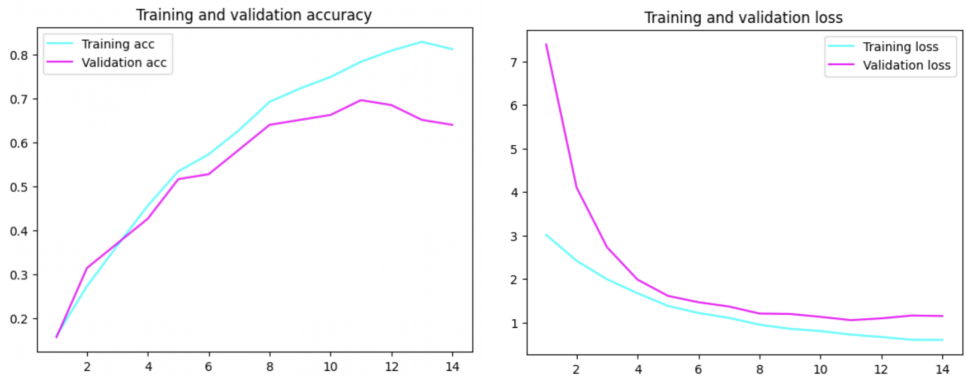


Figure A7: Training and validation accuracy and loss for the model trained with joint features, utilizing a random seed of 9.

D Confusion Matrices

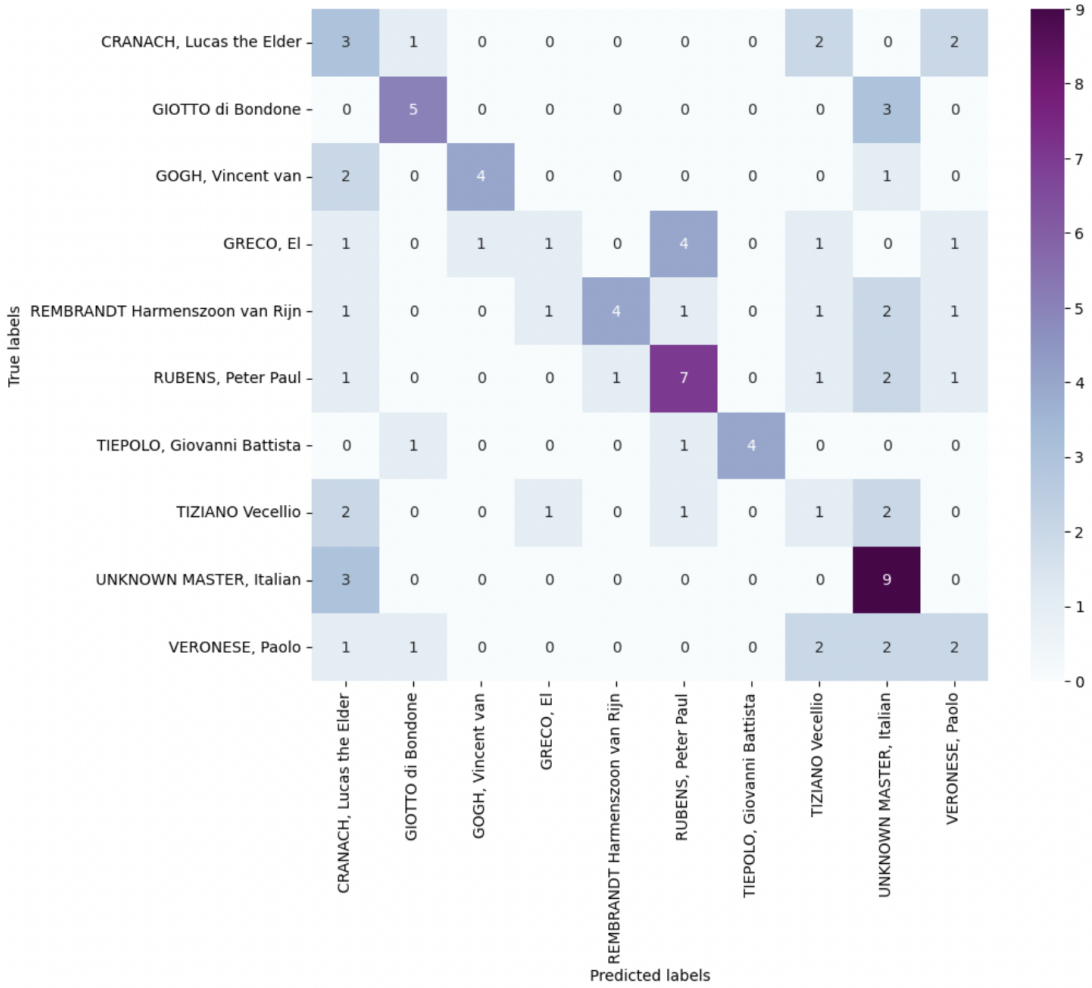


Figure A8: Confusion matrix for the model trained with visual features, utilizing a random seed of 9.

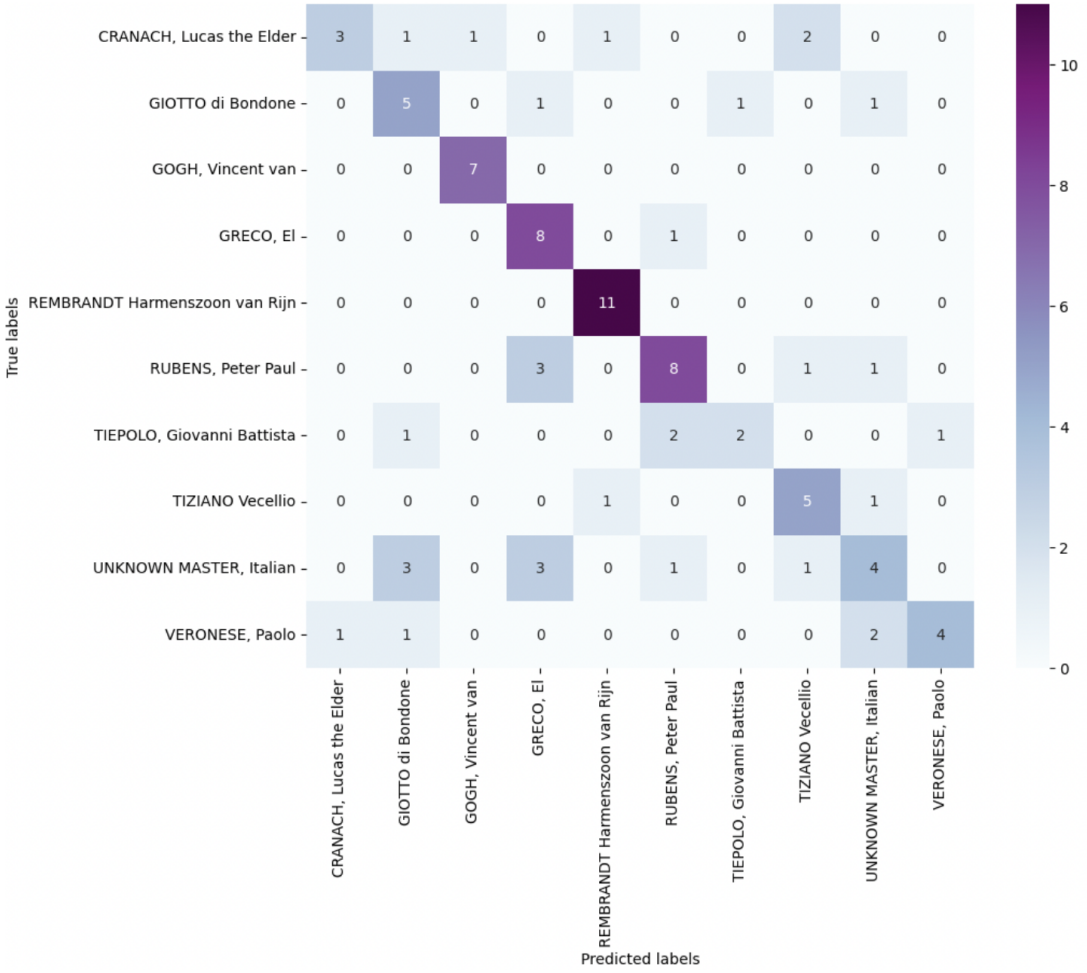


Figure A9: Confusion matrix for the model trained with joint features, utilizing a random seed of 9.