

Harnessing Self-Supervised Features for Art Classification

Federico Melis^{1,†}, Davide Bilardello^{1,†}, Emanuele Prato^{1,†}, Evelyn Turri¹ and Lorenzo Baraldi¹

¹University of Modena and Reggio Emilia, Modena, Italy

Abstract

Classifying artworks presents a significant challenge due to the complex interplay of fine-grained details and abstract features that condition the style or genre of an artwork. This paper presents a systematic investigation of the effectiveness of supervised and self-supervised backbones as feature extractors for both artwork classification and retrieval, with a particular focus on paintings. We conduct an extensive experimental evaluation using the DINO family and CLIP models, assessing multiple classification strategies and feature representations. Our results demonstrate that employing a self-supervised backbone leads to consistent improvements in artwork classification performance. Moreover, our work provides insights into the applicability of classification and retrieval modules in real-world applications, such as virtual reality (VR) applications that support museum navigation.

1. Introduction

Museums are increasingly adopting digital technologies to enrich visitor engagement with cultural heritage. In particular, Augmented Reality (AR) and Virtual Reality (VR) applications have emerged as significant enhancers for the overall experience in museums. These applications rely on structured digital information to support interactive narratives and personalized navigation. In museums centered on static artworks, such as paintings, having an immersive interaction becomes challenging, and AR applications often depend primarily on their metadata. Accurate style and genre classification is a crucial task in the field, facilitating thematic exploration and recommendations of related artworks based on visual similarity.

Recent advances in deep learning methodologies have demonstrated remarkable performance across various domains. In particular, Convolutional Neural Network (CNN) architectures [1, 2, 3] have been highly successful, advancing research in tasks such as object detection and classification, making them more efficient and reliable. However, tasks such as artistic style or genre recognition pose additional challenges: labels are subjective, class boundaries are subtle, and datasets are long-tailed. These factors limit the effectiveness of a fully supervised approach, making it prone to overfitting and reducing its generalization capability. Self-supervised learning (SSL) is particularly effective in disentangling the training objective from the downstream task, capturing more transferable visual structures, and providing robust feature representations. To address the challenge of style and genre classification for artworks, where the required classification is highly specific, we leverage DINO [4, 5, 6] and CLIP [7] as feature extractors, enabling the capture of general and subtle visual patterns that are less dependent on task-specific labels.

This study presents a comprehensive analysis of self-supervised visual representations for artistic style recognition. Specifically, we examine pre-trained vision encoders whose training objectives are decoupled from the downstream task, such as DINO [4, 5, 6] and CLIP [7]. We analyze several models on different strategies, *i.e.* zero-shot inference, KNN zero-shot, and linear classification, highlighting the importance of decoupling feature extraction from the classification stage, and providing a clearer understanding of how this design choice affects stylistic classification performance.

Furthermore, these components can be easily integrated into real-world applications. For instance, both classification and retrieval components can support AR-based museum navigation systems and

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

[†] These authors contributed equally.

✉ 287301@studenti.unimore.it (F. Melis); 285039@studenti.unimore.it (D. Bilardello); 284215@studenti.unimore.it (E. Prato); evelyn.turri@unimore.it (E. Turri); lorenzo.baraldi@unimore.it (L. Baraldi)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

curator tools, assisting museum professionals in categorizing previously unlabeled paintings by style and genre.

We summarize the main contributions of this work as follows:

- We demonstrate the effectiveness of self-supervised feature extractors for artistic classification on the WikiArt dataset [8], highlighting the critical role of decoupled features.
- We provide a systematic evaluation of three different classification strategies for self-supervised models, including a comparison with EfficientNetV2 [9], used as a supervised baseline.
- We perform a retrieval study that highlights the quality of the vision features of our best self-supervised model for the task of artistic style and genre recognition.

2. Related Work

WikiArt and Artistic Classification. The WikiArt [8] dataset offers a diverse benchmark for investigating the classification of artistic attributes, including style, genre, and artist. Early studies [10, 11, 12] highlight that the wide range of artistic expression poses a significant challenge for traditional computer vision methods. More recent approaches leverage deep learning [13, 14], achieving improved performance in stylistic classification and offering interesting insights into how visual representations can bridge the gap between human-level understanding of art while capturing subtle artistic details. Despite these advances, classifying artworks by style and genre remains a persistent challenge, largely due to the inherently abstract and subjective nature of these concepts. Most existing works focus on fine-tuning models pre-trained on large-scale datasets and adapting the extracted features to the artistic domain. Cetinic *et al.* [13] propose a fine-tuning stage on a convolutional neural network for fine art classification. In contrast, the application of self-supervised and weakly-supervised learning for art classification leaves room for further research. CLIP-Art [15] proposes to use CLIP [7] vision features in a zero-shot manner for artistic classification, leveraging natural language information after fine-tuning on the iMet dataset [16]. However, it remains unclear how self-supervised features transfer knowledge to stylistic classification. In this work, we address this gap by investigating performance differences across a wide range of models.

Self-Supervised Learning. The increasing dimension of models and the discoveries in the field of scaling laws [17] have amplified the demand for large-scale datasets. However, in the vision domain, annotating such datasets is often prohibitively expensive or even infeasible. Consequently, significant research has focused on developing methods that achieve strong performance without relying on human-labeled data. Self-supervised learning aims to train a model without explicit training on labeled targets. Weakly-supervised approach, such as CLIP [7], leveraging contrastive learning to operate in a high-dimensional space by minimizing the distance between positive pairs while maximizing it for negative pairs. On the other hand, self-supervised methods such as BYOL [18] employ a siamese architecture in which an online network predicts the representation produced by a target network for a different augmented view of the same image. Similarly, DINO [4] adopts a student-teacher self-distillation paradigm to align representations for different augmentations. Features extracted from DINO models have proven to be effective across multiple downstream tasks. Pérez-García *et al.* [19] show comparable or better performance than biomedical-language supervised models by attaching task-specific decoder heads to a pre-trained DINOv2 [5] backbone. Several works also exploit only the visual representations of CLIP, omitting the textual component. For example, CLIP Fusion [20] leverages multi-scale semantic features from the pre-trained CLIP visual backbone, in order to define a novel metric for detecting video frame interpolation artifacts. Wu *et al.* [21] adapt the CLIP vision encoder for few-shot classification by training an adapter on the original CLIP visual-text embeddings with an additional MLP classifier, combining similarity and classification losses to enhance class separation with limited data.

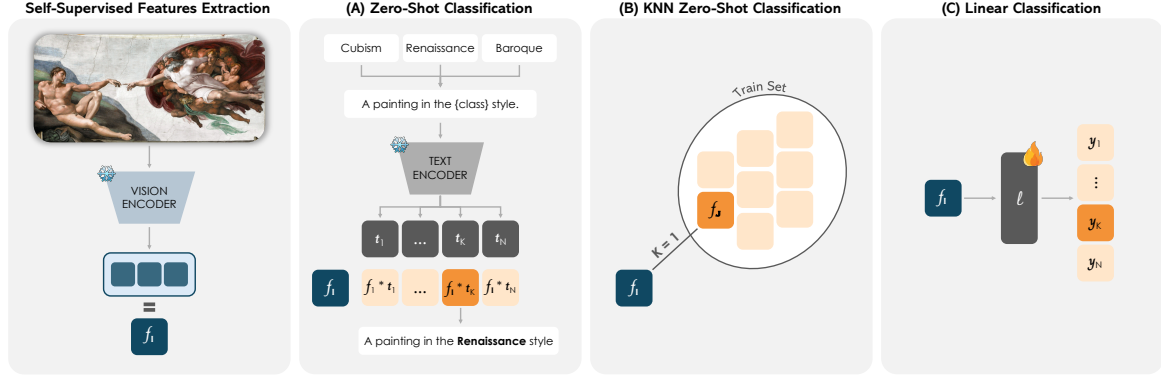


Figure 1: The figure shows the self-supervised feature extraction phase and each of the proposed classification methods: (A) Zero-Shot Classification, (B) KNN Zero-Shot Classification and (C) Linear Classification.

3. Approach

3.1. Task Description

We address two related artistic recognition tasks: artistic style classification and artistic genre classification. Style classification assigns each painting to a specific art style, reflecting characteristic visual conventions, compositional patterns, and historical influences, whereas genre classification emphasizes the thematic content of the artwork, such as landscape, portrait, or illustration. These tasks differ fundamentally from classical image classification, where class labels correspond to well-defined objects or concepts explicitly visible in the image. In contrast, artistic style and genre require the extraction of more abstract features that capture underlying artistic movements or thematic categories. Our experiments provide a comprehensive overview of how SSL models with different representational capabilities adapt to these unconventional and highly nuanced classification challenges.

3.2. Features Extractors

We divide the feature extractors analyzed in the work into two groups, *i.e.* supervised and self-supervised. In both cases, we denote the input image as I , the visual backbone as $\mathcal{V}$, and, where applicable, the text encoder as $\mathcal{T}$. The visual encoder produces features denoted by f , while the text encoder generates features denoted by t . Finally, the label set is denoted as $\mathcal{Y} = (y_1, \dots, y_N)$ where N indicates the number of possible classes.

Supervised Baseline. We employ an EfficientNetV2 [9] as a reference for the supervised-based model. Unlike the self-supervised models used for the downstream tasks, this baseline is trained end-to-end specifically for the target task. A linear classification layer is attached to the final pooled embedding and trained separately for the specific tasks of style and genre classification. The visual backbone is pre-trained on ImageNet [22], and jointly fine-tuned during the training of the classification head. This setup reflects a standard supervised transfer-learning approach, allowing the network to adapt its internal knowledge to the artistic domain. The resulting model serves as a benchmark for evaluating the performance of self-supervised features under different downstream strategies.

Self-Supervised Features Extractors. For our self-supervised pipeline, we employ a pre-trained vision encoder from CLIP [7], and DINO [4, 5, 6]. These models are known to have semantically rich and transferable embeddings, thanks to their training strategy on large datasets. To perform our tasks, we use the classification token (CLS) as the extracted feature representation, following common practice in the literature on vision transformer models for classification tasks. We denote as f_I the representation features in output by $\mathcal{V}(I)$. Using a frozen encoder and the CLS single global representation ensures generalization for the two classification tasks, enabling comparison across classification strategies.

3.3. Classification Strategies

We implement three distinct approaches for classification using representations extracted from the self-supervised feature extractors. In detail, two training-free approaches, *i.e.* zero-shot and zero-shot KNN classification, and a linear classification technique. The zero-shot KNN and linear classification methods rely solely on the visual backbone $\mathcal{V}$, whereas the zero-shot method leverages both the visual encoder $\mathcal{V}$ and the text encoder $\mathcal{T}$. Figure 1 illustrates the feature extractor pipeline along with the employed classification strategies: (A) Zero-Shot Classification, (B) KNN Zero-Shot Classification and (C) Linear Classification.

Zero-Shot Classification. The first approach is zero-shot classification, which evaluates the model’s ability to assign labels to images without task-specific training. In particular, given an image I , a label set $\mathcal{Y} = (y_1, \dots, y_N)$ with N the number of classes, a visual encoder $\mathcal{V}$, and a text encoder $\mathcal{T}$, zero-shot classification proceeds by first extracting visual and text features as:

$$f_I = \mathcal{V}(I) \quad \text{and} \quad t_k = \mathcal{T}(p_k) \quad \forall k \in \{1, \dots, N\} \quad (1)$$

where p_k is the textual prompt corresponding to label y_k , and t_k is the resulting text feature vector. The predicted class $\hat{y}$ is then determined as the label whose textual embedding has the highest cosine similarity with the visual feature f_I .

Zero-Shot KNN Classification. The second classification approach is zero-shot k-Nearest Neighbors (KNN), which similarly requires no training or fine-tuning. This method assigns to a query image $I \in \mathcal{D}_{Query}$ the class of the most similar image within a reference set $\mathcal{D}_{Ref}$. Given the reference set $\mathcal{D}_{Ref}$ and the query set $\mathcal{D}_{Query}$, we extract visual features for the query image I and for each reference image $j \in \mathcal{D}_{Ref}$ as

$$f_I = \mathcal{V}(I) \quad \text{and} \quad f_j = \mathcal{V}(j) \quad \forall j \in \mathcal{D}_{Ref}. \quad (2)$$

The predicted class $\hat{y}$ is then assigned as the label of the reference embedding whose feature vector exhibits the highest cosine similarity with the query embedding f_I .

Linear Classification. The third classification approach introduces a trainable linear layer ℓ applied to the frozen pre-trained visual backbone $\mathcal{V}$, enabling the model to learn a classification mapping from the generalized image representation produced by $\mathcal{V}$. Given the feature vector f_I defined in Equation 1, the linear classifier outputs a vector $z \in \mathbb{R}^N$, where the predicted class $\hat{y}$ corresponds to the index of the maximum value in z .

3.4. Retrieval

To obtain a qualitative assessment of the self-supervised feature representations, we implement a retrieval module based on cosine similarity. Visual embeddings for all images in the dataset are extracted using the visual encoder $\mathcal{V}$ as defined in Equation 1, and these embeddings are indexed using FAISS [23]. Given a query image I with its corresponding embedding f_I , the system retrieves the top- K most similar images from the index.

4. Experiments

4.1. Dataset

We utilize the WikiArt dataset [8], a collection of approximately 80,000 paintings, encompassing 27 style classes and 11 genre classes. Examples of styles include Impressionism, Cubism, Baroque, and Romanticism, while genres cover categories such as illustration, abstract painting, landscape, and portrait. The genre taxonomy also includes the “Unknown Genre” label, which we remove from all

Model	Trainable	Style				Genre			
		P	R	F1	acc@1	P	R	F1	acc@1
EfficientNetV2	✓	68.9	68.6	68.3	68.6	82.0	82.2	82.0	82.2
DINO Zero-Shot	✗	29.1	34.8	26.1	27.4	66.9	67.5	63.7	67.2
CLIP Zero-Shot	✗	39.3	41.2	36.8	41.9	69.4	65.6	60.2	64.4
DINO-KNN	✗	63.5	62.4	61.7	63.2	78.5	78.0	78.1	80.2
CLIP-KNN	✗	69.2	68.6	68.3	70.6	80.3	80.9	80.5	81.7
DINO-Linear	✓	71.5	61.4	65.0	65.0	81.6	80.6	81.0	83.6
CLIP-Linear (Ours)	✓	74.3	67.2	69.8	69.2	83.2	82.5	82.8	84.9

Table 1

Comparison of the baseline EfficientNetV2 [9], compared to CLIP-ViT-L/14 [7] and DINOv3-ViT-L/16 [6] with each classification strategy: Zero-Shot, KNN Zero-Shot, and Linear classification. The table reports precision (P), recall (R), F1-score (F1), and top-1 accuracy (acc@1) for both style and genre prediction, and highlights whether additional trainable components are used. All the results are computed on the test set.

the experiments. The dataset is partitioned into train (80%), validation (10%), and test (10%) splits. The validation set is used exclusively for baseline and for the linear classifier training, while train and test sets are fixed across all the experiments.

4.2. Implementation Details

We detail the experimental setup for the three classification strategies in the following paragraphs to facilitate reproducibility. We report hardware specifications, the hyperparameters used, and a concise summary of methodological details. All experiments were conducted on a single NVIDIA A40 (48 GB). The random seed is fixed to 42 for reproducibility.

Supervised Baseline. We employ two separate visual backbones, one for the style classification task and the other for genre classification, both initialized from identical EfficientNetV2-L checkpoints [9] pre-trained on ImageNet [22]. The default final MLP layer of EfficientNetV2-L is replaced with a linear layer mapping the model’s embedding dimension to the respective number of classes. We apply a dropout [24] rate of 0.4 before the linear layer for regularization. Models are trained for 10 epochs with a batch size of 16, and the best-performing model is selected for evaluation. The full EfficientNetV2-L backbone is fine-tuned using cross-entropy loss with a learning rate of 10^{-5} and a weight decay of 10^{-5} [25], while the classification head is trained with the same loss but a higher learning rate of 10^{-4} . All optimization is performed using the Adam optimizer [26].

Self-Supervised Features Extractors. We employ two self-supervised feature extractors: CLIP-ViT-L/14 and DINOv3-ViT-L/16. For the zero-shot classification strategy, CLIP uses its native text encoder, while DINO utilizes the `dino.txt` [27], text encoder specifically aligned with the DINO’s visual encoder representation space. Prompts for style and genre classification are constructed as follows: “A *painting* in the *<class>* style.” for style, and “A *<genre>* painting.” for genre. For zero-shot KNN classification, the reference set $\mathcal{D}_{Ref}$ corresponds to the WikiArt training set, and the query set $\mathcal{D}_{Query}$ corresponds to the test set. Linear classification layers are trained using cross-entropy loss, a learning rate of 10^{-4} , weight decay equal to 10^{-4} , and a batch size of 1024. Training proceeds for up to 100 epochs with early stopping (patience = 5).

In all experiments, we refer to models using the zero-shot approach as CLIP Zero-Shot and DINO Zero-Shot, models evaluated with the KNN strategy are denoted as CLIP-KNN and DINO-KNN, and models with a trainable linear layer as CLIP-Linear and DINO-Linear.

IMAGE	MODEL	STYLE	GENRE	IMAGE	MODEL	STYLE	GENRE
	EfficientNetv2	Impressionism	Landscape		EfficientNetv2	Abstract Expressionism	Abstract Painting
	DINO-Linear	Impressionism	Landscape		DINO-Linear	Abstract Expressionism	Abstract Painting
	CLIP-Linear (Ours)	Impressionism	Landscape		CLIP-Linear (Ours)	Action Painting	Action Painting
	WikiArt	Impressionism	Landscape		WikiArt	Action Painting	Action Painting
	EfficientNetv2	Baroque	Sketch&Study		EfficientNetv2	Impressionism	Genre Painting
	DINO-Linear	Baroque	Sketch&Study		DINO-Linear	Impressionism	Genre Painting
	CLIP-Linear (Ours)	Baroque	Sketch&Study		CLIP-Linear (Ours)	Impressionism	Nude Painting
	WikiArt	Baroque	Sketch&Study		WikiArt	Impressionism	Nude Painting
	EfficientNetv2	Impressionism	Cityscapes		EfficientNetv2	High Renaissance	Portrait
	DINO-Linear	Impressionism	Cityscapes		DINO-Linear	High Renaissance	Abstract Painting
	CLIP-Linear (Ours)	Impressionism	Cityscapes		CLIP-Linear (Ours)	Mannerism Late Renaissance	Abstract Painting
	WikiArt	Impressionism	Cityscapes		WikiArt	Mannerism Late Renaissance	Abstract Painting

Figure 2: Qualitative results for style and genre classification. For each image, we present predictions for EfficientNetV2, DINO-Linear and CLIP-Linear (Ours) compared to WikiArt [8] Ground Truth.

4.3. Performance on WikiArt

Classification. We report in Table 1 the results achieved under the best configuration for each method, explicitly indicating whether additional trainable components are introduced. The finetuned EfficientNetV2-L with a newly trained classification head achieves solid overall performance on both style and genre classification. Notably, DINO-KNN reaches comparable performance to the finetuned EfficientNetV2-L, despite requiring neither additional trainable parameters nor any finetuning phase. Linear classification settings, instead, outperform all the other configurations, benefiting from high-quality visual features combined with a supervised classification head. Although DINO features are generally recognized for capturing rich semantic information, our experiments reveal consistently stronger performance from CLIP across all evaluation protocols. This suggests that style and genre classification rely more heavily on nuanced visual features than on broad semantic structure. In particular, CLIP-KNN outperforms the baseline across all style metrics, whereas the corresponding DINO-KNN model does not. Best performance is then achieved with CLIP-Linear, indicating that features extracted from the frozen backbone are rich and highly discriminative for both tasks, and that a simple linear layer can easily map those features into each class. It is also noteworthy that zero-shot setting, both for DINO and CLIP models, performs poorly, implying that textual prompts provide limited benefit for style and genre recognition, and that these tasks are primarily driven by visual rather than textual alignment.

The qualitative results in Figure 2 illustrate six examples comparing Ground Truth, the baseline, DINO-Linear, and CLIP-Linear. The experiments include the strongest configuration for each self-supervised feature extractor. In the three examples on the left, both baseline, DINO-Linear, and CLIP-Linear correctly predict both style and genre classes. In the images on the right, instead, only CLIP-Linear correctly predicts style and genre, while both the baseline and DINO-Linear exhibit failures on either style or genre classification.

Retrieval. To further assess the semantic richness and discriminative power of the representations extracted by our strongest self-supervised model, CLIP, we conduct a series of qualitative retrieval experiments on the WikiArt [8] dataset. Figure 3 shows three different qualitative examples for retrieval. In details for each row, the image on the left serves as the query for the retrieval stage, while the numbered pictures are the ones retrieved using the strategy explained in Section 3.3. In all three examples, the retrieved images are coherent with the query. In particular, in all three cases, both style and genre are similar to the main picture. In addition, it is clear that the subject is also coherent with the reference. The final row offers a particularly clear illustration: the retrieved paintings not only share the same stylistic and genre characteristics but also exhibit strongly similar compositional features, such as the pose, hair, beard, and glasses of the subject. To highlight the robustness and generality of



Figure 3: This figure shows qualitative retrieval results obtained with CLIP-ViT-L/14. The query images are depicted in the first column for each row, and they span across different classes of style and genre. For each query sample, we show top-5 best retrieved images.

these features, we include an additional retrieval example using a query image outside the WikiArt dataset. The query depicts "*La creazione di Adamo*" by Michelangelo. Remarkably, the top retrieved image from WikiArt corresponds to the same artwork, despite clear differences in color palette and rendering. CLIP successfully identifies the match by relying on deeper visual and semantic cues rather than superficial appearance. The remaining retrieved images also maintain strong consistency with the query in both style and genre, mirroring the behavior observed in the other examples.

5. Conclusion

In this work, we present a systematic overview of self-supervised learning backbones applied to the challenging task of classifying artistic style and genre. By benchmarking diverse approaches, we demonstrate that leveraging pre-trained features from DINO and CLIP significantly benefits artwork classification. Our experiments highlight the value of decoupling feature extraction from the downstream classification task. Most surprisingly, this held true for the CLIP-Linear setup, which performed better than full fine-tuning of the EfficientNetV2 baseline for both style and genre classification. This suggests that multi-modal pre-training in CLIP captures more abstract artistic aspects compared to standard visual supervision. Moreover, the zero-shot KNN strategy proves to be remarkably competitive, achieving performance comparable to supervised methods without any additional training. This further validates the intrinsic semantic quality of self-supervised visual features verified by our qualitative retrieval analysis.

These findings have direct implications for the cultural heritage sector. The explored architectures and strategies are particularly well-suited for integration into immersive AR and VR museum experiences, enabling context-aware navigation and dynamic artwork recommendation for visitors. Additionally, the use of pre-trained features opens opportunities for scalable applications in large-scale digital archives, where manual labeling is often infeasible, supporting more adaptive organization of art collections. The approach also has the potential to enhance accessibility, for instance by providing visually enriched navigation systems for audiences with diverse needs. Finally, it offers educational value, allowing students and learners to explore art collections by style or genre, thereby improving engagement and understanding of artistic heritage.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

- [1] 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.
- [2] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations (ICLR 2015), 2015.
- [3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
- [4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9650–9660.
- [5] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).
- [6] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, et al., Dinov3, arXiv preprint arXiv:2508.10104 (2025).
- [7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.
- [8] W. R. Tan, C. S. Chan, H. E. Aguirre, K. Tanaka, Improved artgan for conditional synthesis of natural image and artwork, IEEE Transactions on Image Processing 28 (2018) 394–409.
- [9] M. Tan, Q. Le, Efficientnetv2: Smaller models and faster training, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 10096–10106. URL: <https://proceedings.mlr.press/v139/tan21a.html>.
- [10] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, H. Winnemoeller, Recognizing image style, arXiv preprint arXiv:1311.3715 (2013).
- [11] R. S. Arora, A. Elgammal, Towards automated classification of fine-art painting style: A comparative study, in: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), IEEE, 2012, pp. 3541–3544.
- [12] J. Zujovic, L. Gandy, S. Friedman, B. Pardo, T. N. Pappas, Classifying paintings by artistic genre: An analysis of features & classifiers, in: 2009 IEEE international workshop on multimedia signal processing, IEEE, 2009, pp. 1–5.
- [13] E. Cetinic, T. Lipic, S. Grgic, Fine-tuning convolutional neural networks for fine art classification, Expert Systems with Applications 114 (2018) 107–118.
- [14] Y. Hong, J. Kim, Art painting detection and identification based on deep learning and image local features, Multimedia Tools and Applications 78 (2019) 6513–6528.
- [15] M. V. Conde, K. Turgutlu, Clip-art: Contrastive pre-training for fine-grained art classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021, pp. 3956–3960.
- [16] C. Zhang, C. Kaeser-Chen, G. Vesom, J. Choi, M. Kessler, S. Belongie, The imet collection 2019 challenge dataset, 2019.
- [17] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, arXiv preprint arXiv:2001.08361 (2020).

- [18] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al., Bootstrap your own latent-a new approach to self-supervised learning, *Advances in neural information processing systems* 33 (2020) 21271–21284.
- [19] F. Pérez-García, H. Sharma, S. Bond-Taylor, K. Bouzid, V. Salvatelli, M. Ilse, S. Bannur, D. C. Castro, A. Schwaighofer, M. P. Lungren, M. T. Wetscherek, N. Codella, S. L. Hyland, J. Alvarez-Valle, O. Oktay, Exploring scalable medical image encoders beyond text supervision, *Nature Machine Intelligence* 7 (2025) 119–130. URL: <https://doi.org/10.1038/s42256-024-00965-w>. doi:10.1038/s42256-024-00965-w.
- [20] G. M. Çökmez, Y. Zhang, C. Schroers, T. O. Aydin, Clip-fusion: A spatio-temporal quality metric for frame interpolation, in: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2025, pp. 7450–7459.
- [21] G. Wu, J. Chen, W. Zhang, R. Wang, Feature adaptation with clip for few-shot classification, in: *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, 2023, pp. 1–7.
- [22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.
- [23] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library, *IEEE Transactions on Big Data* (2025).
- [24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, *The journal of machine learning research* 15 (2014) 1929–1958.
- [25] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: *International Conference on Learning Representations*, 2019. URL: <https://openreview.net/forum?id=Bkg6RiCqY7>.
- [26] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, *CoRR abs/1412.6980* (2014). URL: <https://api.semanticscholar.org/CorpusID:6628106>.
- [27] C. Jose, T. Moutakanni, D. Kang, F. Baldassarre, T. Darcet, H. Xu, D. Li, M. Szafraniec, M. Ramamonjisoa, M. Oquab, O. Siméoni, H. V. Vo, P. Labatut, P. Bojanowski, Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 24905–24916.