

# DFA-CON: A CONTRASTIVE LEARNING APPROACH FOR DETECTING COPYRIGHT INFRINGEMENT IN DEEFAKE ART

Haroon Wahab\* Hassan Ugail\* Irfan Mehmood†

\*School of Computer Science, AI and Electronics, University of Bradford, UK

†School of Management, University of Bradford, UK

## ABSTRACT

Recent proliferation of generative AI tools for visual content creation—particularly in the context of visual artworks—has raised serious concerns about copyright infringement and forgery. The large-scale datasets used to train these models often contain a mixture of copyrighted and non-copyrighted artworks. Given the tendency of generative models to memorize training patterns, they are susceptible to varying degrees of copyright violation. Building on the recently proposed DeepfakeArt Challenge benchmark, this work introduces DFA-CON, a contrastive learning framework designed to detect copyright-infringing or forged AI-generated art. DFA-CON learns a discriminative representation space, posing affinity among original artworks and their forged counterparts within a contrastive learning framework. The model is trained across multiple attack types, including inpainting, style transfer, adversarial perturbation, and cutmix. Evaluation results demonstrate robust detection performance across most attack types, outperforming recent pretrained foundation models. Code and model checkpoints will be released publicly upon acceptance.

**Index Terms**— Deepfake, Art forgery, Contrastive learning, Generative AI

## 1. INTRODUCTION

The growing availability of generative AI tools for visual content creation has raised critical concerns around copyright infringement, especially in the domain of visual artworks [1]. Generative models trained on large-scale, web-scraped datasets often absorb patterns from both copyrighted and public domain images, making them prone to reproducing unauthorized content [2]. This phenomenon is particularly problematic in the context of AI-generated art, where stylistic and structural similarities to original artworks may constitute legal or ethical violations [3]. Given the ease with which forged or derivative artworks can be created and distributed, there is a compelling need for automated methods to assess the originality and attribution of generative outputs.

To study and benchmark the detection of AI-generated

art forgeries, the recently proposed DeepfakeArt Challenge [4] provides a comprehensive dataset comprising over 32,000 image pairs spanning a variety of generative manipulation techniques. Each entry in the dataset consists of a pair of images—either a forged/generated version of an original artwork or two dissimilar, unrelated images. The manipulated images cover several attack types including inpainting, style transfer, adversarial perturbation, and cutmix, simulating realistic scenarios of content misuse. This dataset enables the development of algorithms that go beyond pixel-level artifact detection, focusing instead on semantic similarity and visual attribution, a crucial capability when detecting copyright violations in generative art.

To address this problem, we propose DFA-CON, a supervised contrastive learning framework designed to detect copyright-infringing or forged art generated by AI models. Rather than treating forgery detection as a traditional classification problem, DFA-CON learns an embedding space that encodes semantic similarity between original and manipulated images. Using supervised contrastive loss, the model is trained to pose affinity among original artworks and their forged counterparts, while distancing unrelated images within the same batch. This representation-level formulation allows for greater generalization across manipulation types and better discrimination of subtle, high-quality forgeries, which are common in the domain of AI-generated art.

Our main contributions are summarized as follows:

- We propose a supervised contrastive learning framework, DFA-CON, for training visual encoders to detect copyright infringement in DeepFake art. To the best of our knowledge, this is the first work that introduces a dedicated model tailored specifically for infringement detection in AI-generated artworks.
- We conduct a comprehensive comparison between DFA-CON and recent general-purpose vision foundation models to assess the need for domain-specific training in the context of generative art forensics.
- We release our model checkpoints and code to facilitate future research. The codebase is modular and supports

integration of any embedding model via a standardized wrapper interface, enabling rapid experimentation and reproducibility.

## 2. RELATED WORKS

### 2.1. AI Generated Art

Recent advancements in generative models such as GANs and diffusion models have enabled the creation of highly realistic and stylistically rich visual artworks [5, 6]. While this has opened new avenues in creative expression, it has also raised significant concerns regarding authorship, originality, and copyright infringement [3]. Prior work in AI-generated art has largely focused on generation techniques, artistic style transfer, and aesthetics modeling, with less attention paid to post-generation verification or attribution. The DeepfakeArt Challenge benchmark [4] was introduced to bridge this gap by facilitating research on forgery and contamination detection in generative art settings.

### 2.2. Deep Fake Detection

Deep Fake detection has emerged as a critical subfield in computer vision, primarily aimed at identifying synthetic or manipulated facial content [7]. Detection methods typically focus on spatial or frequency domain inconsistencies, fine-grained artifact analysis, or classification of pixel-level distortions. However, these methods are often limited to human faces and fail to generalize to non-photorealistic domains like art. In contrast, the task of detecting forgeries in AI-generated artworks involves semantic similarity and contextual interpretation rather than artifact spotting, necessitating representation-based methods like ours.

### 2.3. Contrastive Learning

Contrastive learning has shown great promise in learning discriminative and generalizable visual representations [8]. In particular, supervised contrastive loss [9] has been used effectively in domains where fine-grained similarity modeling is essential. Unlike self-supervised contrastive learning, which relies on augmentations of the same image, supervised variants utilize structured label information to align semantically similar samples. Our method builds on this paradigm by using original-forged image pairs as positives and treating dissimilar pairs as implicit negatives, resulting in an embedding space optimized for semantic attribution and forgery detection.

## 3. PRELIMINARIES

### 3.1. Copyright Infringement in Art

Deep Fake art generative models are vulnerable to violating copyright terms by producing images that mimic or closely resemble content protected under copyright [10]. A formal mathematical formulation of copyright infringement in this context is introduced in [4]. For clarity and contextual relevance, we present a simplified version here:

$$\|A(y)_\Omega - A(T(\hat{x}))_\Omega\| < f(|\Omega|) \cdot \delta \quad (1)$$

Here,  $y$  denotes the generated image and  $\hat{x}$  is a potentially copyright-protected training image. The operator  $T(\cdot)$  represents a geometric transformation (e.g., resizing, flipping, rotation), and  $\Omega$  is a region of significant size within the image. The function  $A(\cdot)$  defines the domain of representation—either raw pixel space or an edge-based representation. The term  $f(|\Omega|)$  is a monotonic function adjusting sensitivity based on region size, and  $\delta$  is a fixed similarity threshold. Infringement is said to occur if the distance in the representation space for any region  $\Omega$  falls below this threshold.

### 3.2. Deepfake Art Challenge

The DeepfakeArt Challenge [4] provides a large-scale benchmark for detecting copyright-infringing or adversarially manipulated images in the domain of AI-generated art. The dataset comprises over 32,000 image pairs, each labeled as either similar (indicating a generated version of an original artwork) or dissimilar (completely unrelated pairs). Each similar pair corresponds to a specific form of generation-based manipulation, known as *attack type*. The challenge enables the development and evaluation of models capable of distinguishing original artworks from manipulated or forged counterparts.

#### 3.2.1. Inpainting

Inpainting attacks simulate localized editing by removing a region of the original image and replacing it using a generative inpainting model such as Stable Diffusion [6]. The masked region is filled in a semantically consistent but potentially infringing way.

#### 3.2.2. Style Transfer

Style transfer-based forgeries are generated by transferring the artistic style of the original artwork onto a different content image, producing a stylized image that may still share strong visual resemblance to the original [11].

### 3.2.3. Adversarial Perturbation

Adversarial examples are crafted by introducing imperceptible pixel-level changes to the original image that lead to semantic drift or visual mimicry.

### 3.2.4. CutMix

CutMix attacks generate forged samples by cutting and pasting regions from multiple images, including the original. This composite manipulation results in a hybrid image that may retain recognizable features from copyrighted content.

**Table 1.** Distribution of Similar Pairs by Attack Type

Attack Type	#Pairs	Percentage
Inpainting	5063	39%
Style Transfer	3074	24%
Adversarial	2730	21%
CutMix	2000	16%
<b>Total</b>	<b>12,867</b>	<b>100%</b>

## 4. METHODOLOGY

This section describes the overall methodology employed in this work. We first present DFA-CON, a contrastive representation learning framework designed to detect copyright infringement in AI-generated art (see Fig. 1). We then introduce an inference-time detection pipeline that leverages the pretrained embedding model to evaluate whether a generated image constitutes a potential infringement (see Fig. 2).

### 4.1. DFA-CON

DFA-CON consists of three core components: (a) forgery-aware contrastive sampling, (b) representation learning using an encoder and projection head, and (c) a supervised contrastive loss objective for training.

#### 4.1.1. Forgery-Aware Contrastive Sampling

In order to adapt the contrastive learning paradigm to the task of copyright-infringement detection in generative art, a sampling strategy is employed that reflects the structural assumptions of the problem as formalized in Section 3.1. Each original artwork in the dataset is treated as an anchor instance, denoted by  $i$ , and its corresponding forgeries—produced via different attack types—are collected into a set of positives  $\mathcal{P}(i)$ . These forged versions represent multiple semantically similar views of the anchor image under the notion of copy-right violation. A batch  $\mathcal{B}$  is formed by sampling multiple such anchors and their associated positives. For each anchor  $i$ , the negative set  $\mathcal{N}(i)$  is implicitly defined as all other instances in the batch that do not belong to  $\{i\} \cup \mathcal{P}(i)$ , i.e.,

$\mathcal{N}(i) = \mathcal{B} \setminus (\{i\} \cup \mathcal{P}(i))$ . This formulation enables batch-wise supervised contrastive training where similarity is learned under the forgery-aware relational structure of the data.

#### 4.1.2. Representation Learning

The proposed framework is structured to learn semantically meaningful representations that support contrastive learning objectives. A two-stage architecture is employed, consisting of a backbone encoder and a projection head. The encoder maps input images from pixel space into a high-dimensional semantic representation space, while the projection head transforms these representations into a lower-dimensional space optimized for contrastive loss. In this work, a ResNet-50 architecture is used as the encoder, where the final fully connected classification layer is removed. This results in a representation vector in  $\mathbb{R}^{2048}$  for each input image.

Two variants of the projection head are explored as part of an ablation study. The first variant is a linear projection head, which maps the encoder output from  $\mathbb{R}^{2048}$  directly to a 128-dimensional space. The second variant is a multilayer perceptron (MLP) that includes a hidden layer of size 2048 with a non-linear activation, followed by a projection to  $\mathbb{R}^{128}$ . These representations are then used for supervised contrastive learning.

#### 4.1.3. Contrastive Loss

To train the model to learn robust and discriminative representations for forgery detection, we adopt the supervised contrastive (SupCon) loss [9]. SupCon is selected due to its improved robustness and its native support for the multi-positive setting, which aligns well with our data formulation, where each anchor may have multiple forged variants. By leveraging all valid positive associations for a given anchor within a batch, the loss encourages consistent representation of similar images and separation from unrelated ones.

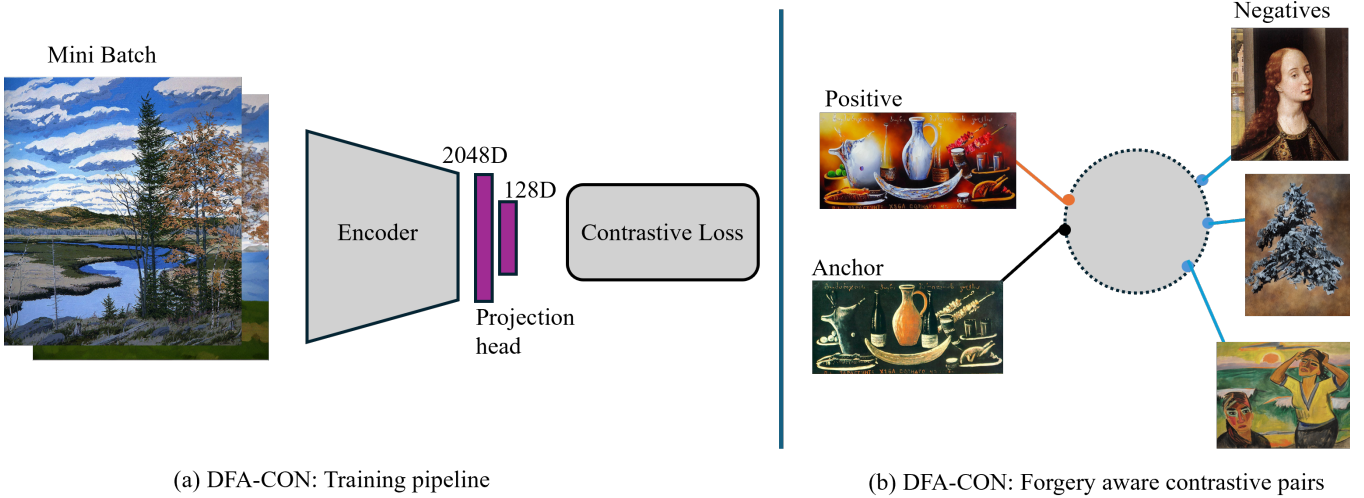
Given a batch  $\mathcal{B}$  of size  $N$ , and a representation  $\mathbf{z}_i$  for each sample  $i$ , the SupCon loss for an anchor  $i$  is defined as:

$$\mathcal{L}_i = \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in \mathcal{B} \setminus \{i\}} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)} \quad (2)$$

Here,  $\mathcal{P}(i)$  denotes the set of positives for anchor  $i$ , and  $\tau$  is a temperature scaling parameter. The final loss is computed by averaging  $\mathcal{L}_i$  over all anchors in the batch. This formulation allows the model to leverage all known positive associations in a batch while contrasting against a shared set of negatives.

#### 4.1.4. Training details

The model is trained using the SupCon loss on mini-batches constructed with the forgery-aware sampling strategy described earlier. An 80-20% split is applied on the official



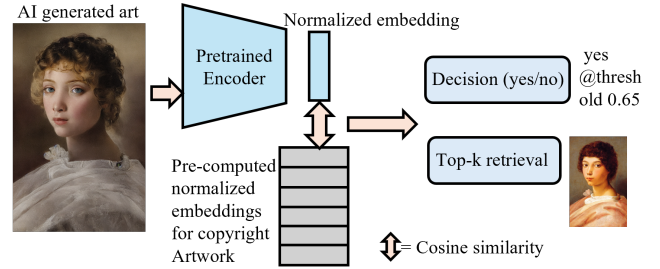
**Fig. 1.** Methodology for training contrastive model to detect copyright infringements in DeepFake Art (DFA-CON)

training split of the dataset ensuring no overlap between original as well as forged versions across the split. All images are resized to  $224 \times 224$  and normalized using ImageNet statistics. The encoder is initialized with ImageNet-pretrained weights, and the entire network is optimized using stochastic gradient descent (SGD) with momentum 0.9. We train the model for 50 epochs, using an initial warm-up phase of 10 epochs followed by cosine annealing of the learning rate. The base learning rate is set to 0.01. Early stopping is applied with a patience of 10 epochs based on validation loss. The temperature parameter  $\tau$  is set to 0.07, following the original SupCon formulation. Training is conducted using a batch size of 128; we also experiment with 32 and 64 as hyperparameters. Both linear and MLP-based projection heads are evaluated as part of an ablation study.

#### 4.2. Copyright Infringement Detection Pipeline

The proposed inference pipeline is designed for practical copyright verification of AI-generated artworks. Consider a scenario where a user possesses a single image or a collection of original, copyright-protected artworks. The goal is to determine whether a given image produced by a generative AI model infringes upon any of the known originals.

As illustrated in Fig. 2, the generated image is first passed through the embedding model, which could be DFA-CON or any other pretrained visual encoder. The resulting embedding is normalized and compared against the set of pre-computed, normalized embeddings of the original artworks using cosine similarity. A threshold-based decision rule is then applied to determine whether the image constitutes a potential infringement. This pipeline is computationally lightweight and scalable, and it also enables top- $k$  retrieval of the most similar original artworks for use in more fine-grained analysis or



**Fig. 2.** Copyright infringement detection pipeline

human-in-the-loop verification.

## 5. EVALUATION

We evaluate our model on an exclusive test split provided by the DeepfakeArt benchmark. A similarity threshold is first determined using validation set and then applied during testing to make binary decisions. If the cosine similarity between the pair in inference exceeds the threshold, the model classifies the pair as similar—indicating a potential copyright violation. Otherwise, the image is considered dissimilar, suggesting no infringement. This formulation naturally aligns with a binary classification setting, and we adopt precision, recall, and F1 score as our primary performance metrics. We report both overall performance across all attack types and per-attack-type results to provide insight into the model’s ro-

**Table 2.** Overall Performance Comparison (Precision, Recall, F1)

Model	P	R	F
ResNet-50 (ImageNet)	0.7988	0.7330	0.7645
ViT-B/16 (ImageNet)	0.813	0.7032	0.7541
DINO-v2 ViT-L/14	0.7181	0.6125	0.6611
CLIP ViT-B/16 (OpenAI)	0.8643	0.7056	0.7769
<b>DFA-CON (Ours)</b>	<b>0.9481</b>	<b>0.7465</b>	<b>0.8353</b>

bustness against specific forms of forgery.

### 5.1. Baseline Models

To assess the effectiveness of DFA-CON, we compare it against several widely used pretrained visual foundation models. To the best of our knowledge, there are no existing models specifically trained for the task of detecting copyright infringement in AI-generated art. Therefore, we consider recent high-capacity embedding models that are commonly used for general-purpose visual representation learning.

The baseline models include: (i) ResNet-50 pretrained on ImageNet [12], (ii) ViT-B/16 pretrained on ImageNet [13], (iii) DINO-v2 ViT-L/14 [14], and (iv) CLIP ViT-B/16 [15]. All baseline models are evaluated using the same inference pipeline described in Section 4.2, where cosine similarity is computed between normalized embeddings and a fixed threshold is used for binary classification.

**Table 3.** Performance by Attack Type (Precision, Recall, F1)

Attack	Model	P	R	F
Inpainting	ResNet-50	0.7378	0.9771	0.8407
	ViT-B/16	0.7634	0.9459	0.8449
	DINO-v2	0.6616	0.6405	0.6509
	CLIP ViT-B/16	0.8137	0.8837	0.8473
	<b>DFA-CON</b>	<b>0.9393</b>	<b>0.9386</b>	<b>0.939</b>
Style Transfer	ResNet-50	0.4955	0.4179	0.4534
	ViT-B/16	0.5	0.4058	0.448
	DINO-v2	0.4566	0.4954	0.4752
	CLIP ViT-B/16	0.6233	0.4802	0.5425
	<b>DFA-CON</b>	<b>0.8923</b>	<b>0.9696</b>	<b>0.9294</b>
Adversarial	ResNet-50	0.6864	0.9954	0.8125
	ViT-B/16	0.7178	1.0	0.8357
	DINO-v2	0.6614	1.0	0.7962
	CLIP ViT-B/16	0.75	0.9977	0.8563
	<b>DFA-CON</b>	<b>0.9168</b>	<b>0.9943</b>	<b>0.9539</b>
CutMix	ResNet-50	0.5906	<b>0.3623</b>	<b>0.4491</b>
	ViT-B/16	0.5626	0.2859	0.3791
	DINO-v2	0.5601	0.2697	0.3641
	CLIP ViT-B/16	<b>0.6347</b>	0.3299	0.4341
	<b>DFA-CON</b>	0.5341	0.0544	0.0987

## 5.2. Performance Comparison

### 5.2.1. Overall Performance

As shown in Table 2, DFA-CON significantly outperforms all baseline foundation models across precision, recall, and F1 score on the overall test set. This suggests that pretrained vision models—despite being trained on large-scale and diverse datasets—do not produce task-aligned representations sufficient for detecting copyright violations in DeepFake art. In contrast, DFA-CON benefits from its supervised contrastive training on explicitly structured forgery data, enabling it to learn more discriminative and attribution-aware embeddings.

The observed performance gap highlights an important limitation of general-purpose visual encoders. While such models are effective at broad semantic understanding, they may not capture the fine-grained visual cues and structural similarities that distinguish authentic artworks from forged variants. This is especially relevant in the context of DeepFake art, where visual mimicry often occurs at a stylistic or compositional level rather than through overt image artifacts. By training on forgery-aware pairings, DFA-CON is able to better internalize the nuanced characteristics of copyright infringement in generative content.

### 5.2.2. Per-Attack-Type Performance

Table 3 presents a breakdown of performance across individual attack types. DFA-CON consistently achieves the highest F1 scores on inpainting, style transfer, and adversarial attacks, confirming its ability to generalize across varied forgery strategies. However, performance notably declines on the CutMix attack, where DFA-CON underperforms compared to all baseline models. This unexpected result may be attributed to the fundamentally different nature of CutMix-based forgeries, which involve compositional splicing of image regions from multiple sources rather than style imitation or localized perturbations. Such hybrid manipulations may introduce ambiguous visual signals that current contrastive supervision struggles to capture effectively. While this remains speculative, it suggests that additional investigation is needed to better understand model behavior on this attack type and to explore whether tailored contrastive sampling or auxiliary supervision could improve performance in such scenarios.

## 6. ABLATION STUDY

We conduct an ablation study to examine the impact of the probe point within DFA-CON, specifically evaluating whether representations extracted from different levels of the model affect detection performance. Results indicate that using embeddings directly from the encoder output in  $\mathbb{R}^{2048}$  yields the highest scores across all metrics. In comparison, probing from the projection head—whether using a linear or

MLP variant projecting to  $\mathbb{R}^{128}$ —leads to a slight degradation in performance, typically in the range of 1–2%. This suggests that while the projection head is effective during training for optimizing the contrastive loss, the encoder-level features are better aligned with the downstream task of infringement detection.

## 7. CONCLUSION

This work presented DFA-CON, a supervised contrastive learning framework for detecting copyright infringement in AI-generated art. Our method leverages forgery-aware sampling and contrastive representation learning to distinguish original artworks from their forged counterparts. Extensive experiments on the DeepfakeArt benchmark demonstrate that DFA-CON significantly outperforms several widely used foundation models. We also analyzed performance across different attack types, highlighting the robustness and limitations of our approach. Ablation results showed that encoder-level representations are most effective for the task. We hope our publicly released code and model will serve as a foundation for future work in generative content forensics.

## 8. REFERENCES

- [1] Harry H. Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru, “Ai art and its impact on artists,” in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, 2023, AIES ’23, p. 363–374, Association for Computing Machinery.
- [2] Gerrit van den Burg and Chris Williams, “On memorization in probabilistic deep generative models,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 27916–27928, 2021.
- [3] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein, “Diffusion art or digital forgery? investigating data replication in diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6048–6058.
- [4] Hossein Aboutalebi, Dayou Mao, Rongqi Fan, Carol Xu, Chris He, and Alexander Wong, “Deepfakeart challenge: A benchmark dataset for generative ai art forgery and data poisoning detection,” *arXiv preprint arXiv:2306.01272*, 2023.
- [5] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone, “Can: Creative adversarial networks, generating” art” by learning about styles and deviating from style norms,” *arXiv preprint arXiv:1706.07068*, 2017.
- [6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.
- [7] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu, “Deepfakebench: A comprehensive benchmark of deepfake detection,” *arXiv preprint arXiv:2307.01426*, 2023.
- [8] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton, “Contrastive representation learning: A framework and review,” *Ieee Access*, vol. 8, pp. 193907–193934, 2020.
- [9] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, “Supervised contrastive learning,” *Advances in neural information processing systems*, vol. 33, pp. 18661–18673, 2020.
- [10] Nicky Kriplani, Minh Pham, Gowthami Somepalli, Chinmay Hegde, and Niv Cohen, “Solidmark: Evaluating image memorization in generative models,” *arXiv preprint arXiv:2503.00592*, 2025.
- [11] Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu, “Artistic style transfer with internal-external learning and contrastive learning,” in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2021, NIPS ’21, Curran Associates Inc.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [14] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al., “Dinov2: Learning robust visual features without supervision,” *arXiv preprint arXiv:2304.07193*, 2023.
- [15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastri, Amanda Askell, Pamela Mishkin, Jack Clark, et al.,

“Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.