Investigating Style Similarity in Diffusion Models

Gowthami Somepalli*[1], Anubhav Gupta*[1], Kamal Gupta[1], Shramay Palta[1], Micah Goldblum[3], Jonas Geiping[2], Abhinav Shrivastava[1], and Tom Goldstein[1]

[1] University of Maryland, College Park
[2] ELLIS Institute, MPI for Intelligent Systems
[3] Columbia University
[4] https://somepago.github.io/csd

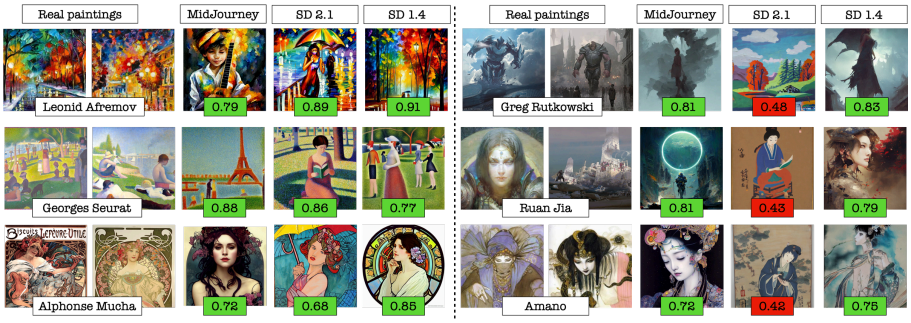

Fig. 1: Original artwork of 6 popular artists and the images generated in the style of these artists by three popular text-to-image generative models. The numbers displayed below each image indicate the similarity of the generated image with the artist's style using the proposed method. A high similarity score suggests a strong presence of the artist's style elements in the generated image. Based on our analyses, we postulate that three artists on the right were removed (or unlearned) from SD 2.1 while they were present in MidJourney and SD 1.4. Please refer to Section 2 for more details.

Abstract. Generative models are now widely used by graphic designers and artists. Prior works have shown that these models remember and often replicate content from their training data during generation. Hence as their proliferation increases, it has become important to perform a database search to determine whether the properties of the image are attributable to specific training data, every time before a generated image is used for professional purposes. Existing tools for this purpose focus on retrieving images of similar *semantic content*. Meanwhile, many artists are concerned with *style* replication in text-to-image models. We present a framework for understanding and extracting style descriptors from images. Our framework comprises a new dataset curated using the

*Equal contribution. Correspondence: gowthami@umd.edu.

insight that style is a subjective property of an image that captures complex yet meaningful interactions of factors including but not limited to colors, textures, shapes, *etc.*We also propose a method to extract style descriptors that can be used to attribute style of a generated image to the images used in the training dataset of a text-to-image model. We showcase promising results in various style retrieval tasks. We also quantitatively and qualitatively analyze style attribution and matching in the Stable Diffusion model.

Keywords: Image Style · Style Similarity · Generative models

1 Introduction

Diffusion-based image generators like Stable Diffusion [49], DALL-E [47] and many others [1, 7, 39, 43] learn artistic styles from massive captioned image datasets that are scraped from across the web [54]. Before a generated image is used for commercial purposes, it is wise to understand its relationship to the training data and the origins of its design elements and style attributes. Discovering and attributing these generated images, typically done with image similarity search, is hence becoming increasingly important. Such dataset attribution serves two purposes. It enables users of generated images to understand potential conflicts, associations, and social connotations that their image may evoke. It also enables artists to assess whether and how generative models are using elements of their work.

Despite a long history of research [60], recovering style from an image is a challenging and open problem in Computer Vision. Many existing retrieval methods [8, 45, 46] for large training datasets focus primarily on matching *semantic* content between a pair of images. Understanding the origin of the *style* present in a generated image, however, is much less well understood. To address this gap, we propose a self-supervised objective for learning style descriptors from images. Standard augmentation-based SSL pipelines (e.g. SimCLR and variants) learn feature vectors that are invariant to a set of augmentations. Typically, these augmentations preserve semantic content and treat style as a nuisance variable. In contrast, we choose augmentations that preserve stylistic attributes (colors, textures, or shapes) while minimizing content. Unfortunately, SSL is not enough, as style is inherently subjective, and therefore a good style extractor should be aligned with human perceptions and definitions of style. For this reason, we curate a style attribution dataset, **ContraStyles**, in which images are associated with the artist that created them.

By training with both SSL and supervised objectives, we create a high-performance model for representing style. Our model, CSD, outperforms other large-scale pre-trained models and prior style retrieval methods on standard datasets. Using CSD, we examine the extent of style replication in the popular open-source text-to-image generative model Stable Diffusion [49], and consider different factors that impact the rate of style replication.

To summarize our contributions, we (1) propose a style attribution dataset ContraStyles, associating images with their styles, (2) introduce a multi-label contrastive learning scheme to extract style descriptors from images and show the efficacy of the scheme by zero-shot evaluation on public domain datasets such as WikiArt and DomainNet (3) We perform a style attribution case study for one of the most popular text-to-image generative models, Stable Diffusion, and propose indicators of how likely an artist's style is to be replicated. Code, artifacts, and dataset links are available at `https://somepago.github.io/csd`.

2 Motivation

We present a case study that shows how style features can be used to interrogate a generative model, and provide utility to either artists or users. We consider the task of analyzing a model's ability to emulate an artist's style, and of attributing the style of an image to an artist. We begin by curating a list of 96 artists, primarily sourced from the WikiArt database, supplemented by a few contemporary artists who are notably popular within the Stable Diffusion community[5]. For each artist, we compute a prototype vector by averaging the embeddings of their paintings using our proposed feature extractor, CSD ViT-L. Next, we generate an image for each artist using Stable Diffusion 2.1 with a prompt in the format `A painting in the style of <artist_name>`. We compute the dot product similarity between each generated image's embedding and the artist's prototype. This process was repeated multiple times for each artist, and we plot mean results in Fig. 2. We refer to this quantity as the General Style Similarity (GSS) score for an artist, as it measures how similar a generated image is to a typical image from that artist while using our style representation model. We also plot an analogous style similarity score, but using "content-constrained" prompts. For instance, one prompt template is `A painting of a woman doing <Y> style of <X>` where `X` is the name of the artist and `Y` is some setting like `reading a book` or `holding a baby` etc. See Sec. 7 for all templates.

Each point in Fig. 2 represents an artist. Notice that GSS scores are highly correlated with content-constrained scores, indicating that our feature vectors represent style more than semantic content. Our findings reveal that SD 2.1 is much more capable of emulating some artists than others. Artists like `Leonid Afremov`, `Georges Seurat` exhibit high style similarity scores, and visual inspection of generated images confirms that indeed their style is emulated by the model (Fig. 1 - Original artwork vs SD 2.1). On the other end of the spectrum, artists such as `Ruan Jia` and `Greg Rutkowski` showed low similarity scores, and likewise the generated images bear little resemblance to the artists' work. Interestingly, after completing this study, we discovered that `Greg Rutkowski`'s work was excluded from the training data for the Stable Diffusion 2.1 model, as reported by [11].

This demonstrates that the Style Similarity score can be used by artists to quantify how well a model emulates their style, or it can be used by users to

[5] `https://supagruen.github.io/StableDiffusion-CheatSheet/`

ascertain whether a generated image contains stylistic elements associated with a particular artist. After a thorough inspection of the generations from 96 artists, we hypothesize that a single-image Style Similarity score below 0.5 indicates the absence of the artist's style, whereas a score above 0.8 strongly indicates its presence.

In Figure 1, we show original artworks for 6 artists, and generations from MidJourney [39], Stable Diffusion 2.1 and Stable Diffusion 1.4 [49] for each of these artists. The 3 artists on the left side have high GSS while the ones on the right side have low GSS. Below each generated image, we display the similarity against the artist's prototype vector. We see high image similarity scores in the Midjourney generations and qualitatively these images look stylistically similar to artists' original artworks. We also see the interesting cases of `Greg Rutkowski`, `Ruan Jia`, and `Amano` whose style is captured by Stable Diffusion 1.4, while being notably absent in Stable Diffusion 2.1. This finding is in line with reports suggesting that some of these artists were removed from the training data of Stable Diffusion 2.1 [11]. Based on this analysis, we postulate that `Ruan Jia`, `Wadim Kashin`, `Anton Fadeev`, `Justin Gerard`, and `Amano` were also either excluded from the training data or post-hoc unlearned/removed from Stable Diffusion 2.1.

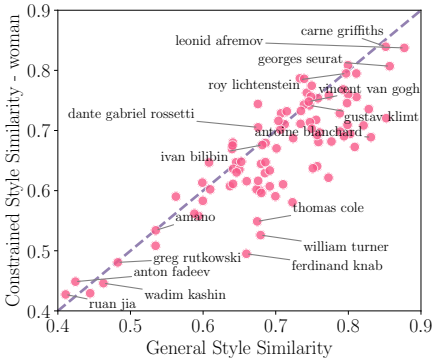

Fig. 2: Style similarity of Stable Diffusion 2.1 generated images against the artist's prototypical representation. On the X-axis, the similarities are depicted when the prompt is not constrained, while the Y-axis represents similarity when the prompt is constrained to generate an image of a "woman" in the artist's style.

3 What is style?

The precise definition of "style" remains in contention, but many named artistic styles (e.g., cubism, impressionism, etc...) are often associated with certain artists. We leverage this social construct, and define style simply as the collection of global characteristics of an image that are identified with an artist or artistic movement. These characteristics encompass various elements such as color usage, brushstroke techniques, composition, and perspective.

Related work. Early computer vision algorithms attempted to model style using low-level visual features like color histograms, texture patterns, edge detection, and shape descriptors. Other computational techniques involve rule-based systems, such as the presence of specific compositional elements, the use of spe-

cific color palettes, or the presence of certain brushstroke patterns to identify specific style characteristics [19, 20, 24, 25, 29, 32, 35, 37, 51, 55, 59, 65, 67].

Modern studies have focused on the task of transferring style from one image to another [13, 17, 22, 34, 41, 62, 68]. Some works have also concentrated on style classification [2, 4, 10, 15, 27, 28, 30, 33, 38, 48, 53]. A limited number of studies address in-the-wild style quantification, matching, and retrieval [14, 23, 36, 50, 66]. In their seminal work, Gatys et al. [17] introduced Gram Matrices as style descriptors and utilized an optimization loop to transfer style. Another approach proposed by Luan et al. [34] includes a photorealism regularization term to prevent distortions in the reconstructed image. Zhang et al. [68] formulated style transfer using Markov random fields. Beyond Gram-based style representation, Chu et al. [10] explored various other types of correlations and demonstrated performance variations.

In a recent work by Lee et al [31], two separate neural network modules were used – one for image style and another for image content – to facilitate image style retrieval. In the most recent related research, Wang et al. [63] developed an attribution model trained on synthetic style pairs, designed to identify stylistically similar images. In contrast to this approach, our method leverages real image pairs, curated automatically through their caption annotations. Despite our training dataset being approx. 16% the size of training data used in [63], we demonstrate that our model significantly outperforms this method on many zero-shot style matching tasks in the later sections.

4 Creating a dataset for style attribution

While many large web datasets now exist, we need one that contains wide variations in artistic styles, and also labels that be used for downstream style retrieval evaluation. Some large-scale datasets specifically designed to handle such a challenge, like BAM [64], are not available in the public domain and others like WikiArt [52] are not large enough to train a good style feature extractor. In the following section, we propose a way to curate a large style dataset out of the LAION [54] Aesthetics 6+ dataset.

ContraStyles: A dataset for style distillation. We curate our own dataset as a subset of LAION [54]. We start off with the 12M image-text pairs with predicted aesthetics scores of 6 or higher. Note that this dataset is extremely unbalanced, with some popular artists appearing much more frequently than others. Also, a large number of images are duplicated within the dataset which is a major issue for the text-to-image models trained on this data [57, 58]. Furthermore, the image captions within the data are often noisy and are often missing a lot of information. We address these challenges and propose a new subset of LAION-Aesthetics consisting of 511,921 images, and 3840 style tags, where each image can have one or more tags. We use this dataset for training our models.

We begin with a bank of styles collated in previous work for image understanding with the CLIP Interrogator [44]. This bank of styles was curated based on typical user prompts for Stable Diffusion. We combine the bank of artists,

mediums, and movement references, to a shortlist of 5600 tags. We then search for these tags in the 12M LAION-Aesthetics captions and shortlist the images that have at least one of the tags present. We further filter out the tags which have over 100,000 hits in the dataset since human inspection found that they refer to common phrases like 'picture' or 'photograph' that do not invoke a distinct style. After discarding images with an unavailable URL, we are left with about 1 million images and 3840 tags. We further deduplicate the images using SSCD [45] with a threshold of 0.8 and merge the tags of images that are near copies of each other. As a by-product, the deduplication also helps with the missing tags in the images, since we can simply merge the text labels of duplicate images. After deduplication, we are left with 511,921 images.

5 Contrastive Style Descriptors (CSD)

Self-Supervised Learning. Many successful approaches to SSL [61] use a contrastive [9] approach, where two views (or augmentations) of the same image in the dataset are sampled and passed to a deep network to get their respective image descriptors. The network is trained to maximize the similarity of two views of the same image and minimize agreement with other images in the batch. Standard choices for augmentations include color jitter, blurring, grayscaling, *etc.*, as these alter the image's visual properties while preserving content. While these are good augmentations for object recognition tasks, they train the network to ignore image attributes associated with style.

Our approach relies on a training pipeline with two parts. First, we use contrastive SSL, but with a set of augmentations that are curated to preserve style. Second, we align our model with human perceptions of style by training on our labelled ContraStyles dataset described in Section 4.

Proposed Approach. We seek a model for extracting image descriptors that contain concise and effective style information. To be a useful, the model should be invariant to semantic content and capable of disentangling multiple styles.

Given a dataset of N labeled images $\{\mathbf{x}_i, l_i\}_{i=1}^{N}$, where each image can have one or more labels from a set of L labels, we define the label vector of the i^{th} image as $l_i = (c_1, c_2, \ldots, c_L)$, where each $c_k \in \{0, 1\}$. As mentioned in the previous section, our multi-label dataset consists of $N = 511,921$ images and $L = 3,840$ style tags. We consider a mini-batch of B images. Each of the images are passed to a Vision Transformer (ViT) [12] backbone, and then projected to a $d-$dimensional vector. We consider two variants of ViT (ViT-B and ViT-L).

Our style descriptors $f_{\text{ViT}}(\mathbf{x}_i) \in \mathbb{R}^d$ are then used to create a matrix of pairwise cosine similarity scores $s_{i,j} = \cos(f_{\text{ViT}}(\mathbf{x}_i), f_{\text{ViT}}(\mathbf{x}_j))$. In order to compute our **multi-label contrastive loss** (MCL), we also compute the groundtruth similarity scores as $\hat{s}_{i,j} = \mathbb{1}\left(l_i^{\text{T}} l_j\right)$, where $\mathbb{1}$ is the indicator function that returns 1 if any of the labels of the images i, j match. Our final loss term reduces to:

$$\mathcal{L}_{\text{MCL}} = -\hat{s}_{i,j} \log \frac{\exp(s_{i,j})/\tau)}{\sum_{k \neq j} \exp(s_{i,k})/\tau)}, \tag{1}$$

where τ is the temperature fixed during the training.

Since our supervised dataset is modest in size, we add a self-supervised objective. We sample two "views" (augmentations) of each image in a batch and add a contrastive SSL term. Standard SSL training routines (e.g., MoCo, SimCLR, BYOL *etc.*) choose augmentations so that each pair of views has the same semantic content, but different style content. These augmentations typically include Resize, Horizontal Flips, Color Jitter, Grayscale, Gaussian Blur, and Solarization [6, 18]. For our purposes, we depart from standard methods by excluding photometric augmentations (Gaussian Blur, Color Jitter), as they alter the style of the image. We keep the following spatial augmentations - Horizontal Flips, Vertical Flips, Resize and Rotation as they keep style intact.

The overall loss function is a simple combination of the multilabel contrastive loss and self-supervised loss $\mathcal{L} = \mathcal{L}_{\text{MCL}} + \lambda \mathcal{L}_{\text{SSL}}$. During inference, we use the final layer embedding and the dot product to compute style similarity between any two images. In our experiments, we found that initializing weights to CLIP [46] ViT-B and ViT-L improves performance.

6 Results

Training details. We present the results for two variants of our model CSD ViT-B and CSD ViT-L version. Both the models are initialized with respective CLIP variant checkpoints and are finetuned for 80k iterations on the ContraStyles dataset on 4 A4000/A5000 GPUs. We use an SGD optimizer with momentum 0.9 and learning rate of 0.003 for the projection layer and $1e-4$ for the backbone. Our mini-batch size per GPU is 16. We use $\lambda = 0.2$ and $\tau = 0.1$ for the final model. The training takes about 8 hours for the base model and around 16 hours for the large model. See the Appendix for more details and ablations.

Task. We perform zero-shot evaluation across multiple datasets on a style-retrieval task. Following [3, 26], we split each dataset into two parts: *Database* and *Query*. Given a query image at test time, we evaluate whether we can find the ground-truth style in its nearest neighbors from the database.

Baselines. We compare our model against a recent style attribution model GDA [63] which is trained via fine-tuning on paired synthetic style data, and VGG [17, 56] Gram Matrices which are often used for neural style transfer applications. Further we compare with CLIP [46] models supervised with free-form text captions, and with other self-supervised models such as DINO [8], MoCo [21], SSCD [45]. We use the embeddings from the last layer for each of these models except for VGG where we use the Gram Matrix [17] of the last

layer. We skipped evaluations of [14, 23, 50] since both pre-trained models and training data are not available.

Metrics. We do nearest neighbour searches for $k \in [1, 10, 100]$, and report Recall@k, mAP@k. We use the standard definitions of these metrics from the retrieval literature. Like [5], we define positive recall as the existence of a correct label in top-N matches and no recall when none of the top-N matches share a label with the query. Similarly, mAP is defined as average over precision at each rank in N, and then averaged over all queries.

Table 1: mAP and Recall metrics on DomainNet and WikiArt datasets. Our model consistently performs the best in all cases except one, against both self-supervised and style attribution baselines.

	DomainNet (mAP@k)			WikiArt (mAP@k)			DomainNet (Recall@k)		WikiArt (Recall@k)		
Method	1	10	100	1	10	100	1	10	1	10	100
VGG Gram [16]	-	-	-	25.9	19.4	11.4	-	-	25.9	52.7	80.4
DINO ViT-B/16 [8]	69.4	68.2	66.2	44.0	33.4	18.9	69.4	93.7	44.0	69.4	88.1
DINO ViT-B/8 [8]	72.2	70.9	69.3	46.9	35.9	20.4	72.2	93.8	46.9	71.0	88.9
SSCD RN-50 [45]	67.6	65.9	62.0	36.0	26.5	14.8	67.6	**95.0**	36.0	62.1	85.4
MOCO ViT-B/16 [21]	71.9	71.1	69.6	44.0	33.2	18.8	72.0	94.0	44.0	69.0	88.0
CLIP ViT-B/16 [46]	73.7	73.0	71.3	52.2	42.0	26.0	73.7	94.5	52.2	78.3	93.5
GDA CLIP ViT-B [63]	62.9	61.6	59.3	25.6	21.0	14.1	62.9	92.3	25.6	56.6	83.8
GDA DINO ViT-B [63]	69.5	68.1	66.1	45.5	34.6	19.7	69.5	93.4	45.5	75.8	89.0
GDA ViT-B [63]	67.1	65.6	64.2	42.6	32.2	18.2	67.1	93.6	42.6	67.6	87.1
CSD ViT-B (Ours)	78.3	77.5	76.0	56.2	46.1	28.7	78.3	94.3	56.2	80.3	93.6
CLIP ViT-L [46]	74.0	73.5	72.2	59.4	48.8	31.5	74.0	94.8	59.4	82.9	95.1
CSD ViT-L (Ours)	**78.3**	**77.8**	**76.5**	**64.56**	**53.82**	**35.65**	**78.3**	94.5	**64.56**	**85.73**	**95.58**

Evaluation Datasets. *DomainNet* [42] consists of an almost equal number of images from six different domains: Clipart, Infograph, Painting, Quickdraw, Real, and Sketch. Upon examination, we observed a strong stylistic resemblance between the Quickdraw and Sketch domains, leading us to exclude Quickdraw from our analysis. The dataset's content information was utilized to categorize the images into two main clusters of content classes. This clustering was achieved through the application of word2vec [40]. The images within the smaller cluster were designated as part of the Query set (20,000 images), and images in the bigger cluster to Database (206768 images). And the second dataset we evaluate all the models is, *WikiArt* [52]. It consists of 80096 fine art images spread across 1119 artists and 27 genres. We randomly split the dataset into 64090 *Database* and 16006 *Query* images. We use the artist as a proxy for the style since there is large visual variation within each genre for them to be considered as independent styles. Under this setting, WikiArt is a challenging retrieval task as the chance probability of successful match is just .09% while it is 20% for DomainNet.

6.1 Analyses and observations

In Table 1, we report the metrics for all the baselines considered and the proposed CSD method using k nearest neighbors on 2 datasets - DomainNet and WikiArt. Note that while style and content are better separated in the case of DomainNet, WikiArt consists of more fine-grained styles and has more practical use cases for style retrieval. Loosely speaking, mAP@k determines what percentage of the nearest neighbors that are correct predictions, while the recall determines what percentage of queries has a correct match in the top-k neighbors. Our model CSD consistently outperformed all the pretrained feature extractors as well as the recent attribution method GDA [63] on both WikiArt and DomainNet evaluations. Note that all models are evaluated in a zero-shot setting. We see the most gains in the WikiArt dataset which is more challenging with chance probability of only 0.09%. When we look at mAP@1, which is same as top-1 accuracy, our base model outperforms the next best model by 5% points on WikiArt and 4.6% points on DomainNet. Our large model out-performs the closest large competitor by similar margins. Given the complexity of the task, these improvements are non-trivial.

We attribute the improvements to a couple of factors, (1) The multi-label style contrastive loss on our curated ContraStyles dataset is quite helpful in teaching the model right styles (2) We hypothesize that these SSL models become invariant to styles because the way they were trained, but we are careful to not strip that away in our SSL loss component by carefully curating non-photometric augmentations in training.

Error Analysis. Even though our model outperforms the previous baselines, our top-1 accuracy for the WikiArt style matching task is still at 64.56. We tried to understand if there is a pattern to these errors. For example, our model is consistently getting confused between impressionist painters `claude monet`, `gustave loiseau`, and `alfred sisley`, all of whom painted many landscapes. They depicted natural scenes, including countryside views, rivers, gardens, and coastal vistas. Another example is `pablo picasso` and `georges braque`, who are both cubist painters. Given the impracticality of analyzing all 1,119 artists in the dataset, we opted for a macroscopic examination by categorizing errors at the art movement level. This approach is visualized in the heatmap presented in Fig. 3. In the heatmap, we see most of the errors concentrated along the diagonal, indicating that while the model often correctly identifies the art movement, it struggles to pinpoint the exact artist. There are instances of off-diagonal errors where the model incorrectly identifies both the artist and their art movement. For example, many Post Impressionism and Realism paintings are assigned to Impressionism artists. Upon closer examination, it becomes apparent that they closely align in terms of historical timeline and geographical origin, both being from Europe. This analysis indicates the nuanced nature of style detection and retrieval in Computer Vision. It suggests that the upper limit for accuracy in this task might be considerably lower than 100%, even for a typical human evaluator, due to the inherent subtleties and complexities involved.

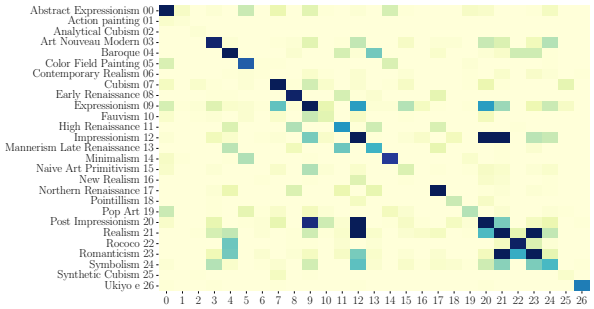

Fig. 3: Confusion Matrix of errors in WikiArt: Art movements are predicted correctly. Errors occur in cases where movements share the same historical timeline and/or are derived from the same earlier movement.

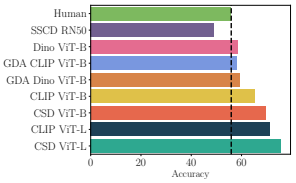

Fig. 4: Human study on Style Retrieval: Turns out **untrained** humans are worse than many feature extractors on matching images from the same artist.

6.2 Human Study

To understand how our models compare to untrained humans, we conducted a small survey on style matching on 30 humans (excluding the authors). Following the convention in other papers [14,28,50,62] and this paper, we assume, 2 images from same artist can be considered stylistic matches. For each query image, we gave 4 answer images out of which only one is from the same artist and hence is the right answer, so chance accuracy is 25%. We used the Artchive dataset introduced in [63] to create this survey and we collected 3 responses per item to break any ties. We present the results in Fig 4. Most interestingly, untrained humans are worse than many feature extractors at this task. SSCD is the only feature extractor that did worse than humans. Our model, CSD outperforms all the baselines on this dataset as well. This underpins the difficulty of style matching and also highlights the superior performance of our feature extractor.

7 Studying style in the wild: Analysis of Stable Diffusion

In the previous section, we have quantitatively shown that our model Contrastive Style Descriptors outperforms many baselines on style matching task. Now we try to address the question, *Can we do style matching on Stable Diffusion generated images?* To answer this question, we first curated multiple synthetic image collections using Stable Diffusion v 2.1 [49] and then compared them against the "ground truth" style matches on **ContraStyles** dataset.

Creating synthetic style dataset. The first challenge in curating synthetic images through prompts is the choice of prompts to be used for the generation. There have been no in-depth quantitative studies of the effect of prompts on generation styles. For this analysis, we chose 3 types of prompts.

Table 2: mAP and Recall of SD 2.1 generated synthetic datasets based on *Simple* prompts and *User-generated* prompts

	Simple prompts (mAP@k)			*User-generated* (mAP@k)			*Simple* prompts (Recall@k)			*User-generated* (Recall@k)		
Method	1	10	100	1	10	100	1	10	100	1	10	100
GDA - DINO	11.6	10.2	7.6	4.45	4.59	4.24	11.6	28.1	52.83	4.45	25.22	67.18
CSD-ViTB	17.53	16.56	12.68	5.85	5.96	5.58	17.53	38.65	61.85	5.85	29.26	74.2
CLIP ViT-L/14	22.3	20.4	16.1	**6.1**	5.7	5.1	22.3	44.5	66.2	**6.1**	26.0	71.7
CSD (Ours)	**24.5**	**23.3**	**18.5**	5.7	**5.9**	**5.6**	**24.5**	**47.2**	**67.5**	5.7	**26.5**	**71.8**

1. *User-generated* prompts: We used a Stable Diffusion Prompts[6] dataset of 80,000 prompts filtered from Lexica.art. We used the test split and then filtered the prompts to make sure at least one of the keywords from the list we curated in Section 4 is present. We then sampled 4000 prompts from this subset for query split generation.
2. *Simple* prompts: We randomly sampled 400 artists which appeared most frequently in user-generated prompts we analysed. We format the prompt as `A painting in the style of <artist-name>`, and we generate 10 images per prompt by varying the initialization seed.
3. *Content-constrained* prompts: We wanted to understand if we can detect style when we constrain the model to generate a particular subject/human in the style of an artist. For this, we used the prompt `A painting of a woman in the style of <artist-name>` or `A painting of a woman reading in the style of <artist-name>` etc., a total of 5 variations per subject repeated two times. We experimented with subjects, `woman`,`dog` and `house` in this study. We provide the exact templates in the appendix.

Table 3: mAP and Recall of SD 2.1 generated synthetic datasets based on *Content-constrained* prompts

	Dog (mAP@k)			House (mAP@k)			Woman (mAP@k)			Dog (Recall@k)			House (Recall@k)			Woman (Recall@k)		
Method	1	10	100	1	10	100	1	10	100	1	10	100	1	10	100	1	10	100
GDA-DINO	2.28	2	1.6	3.9	3.6	2.7	2.2	2.3	2	2.28	8.68	28.73	3.9	12.3	32	2.2	10.1	28.9
CSD-ViTB	4.5	4.31	3.61	4.6	4.29	3.87	7.55	7.83	6.42	4.5	14.36	34.88	4.6	15.03	39	7.55	20.1	42.46
CLIP ViT-L/14	2.3	2.2	1.9	4.5	4.2	3.6	7.4	7.1	6.2	2.2	9.8	29.9	4.5	13.8	35.3	7.4	19.0	41.6
CSD (Ours)	**4.9**	**4.8**	**4.2**	**6.4**	**6.2**	**5.4**	**10.8**	**10.1**	**8.6**	**4.9**	**14.5**	**34.5**	**6.4**	**17.8**	**40.6**	**10.8**	**23.4**	**44.2**

We generate 4000 images for each prompt setting using Stable Diffusion 2.1. There is only one style keyword in *simple* and *content-constrained* prompts, which we also use as a ground truth label for matching tasks. However, *user-generated* prompts can have multiple style labels within the caption, and we consider all of them as ground-truth labels.

[6] `https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts`

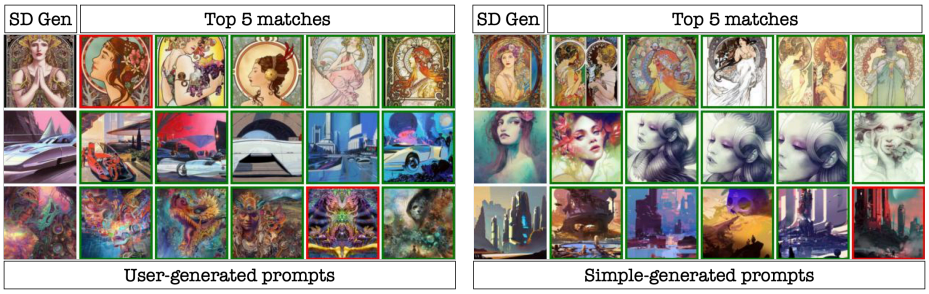

Fig. 5: Nearest "style" neighbors. For each generated image (referred to as SD Gen), we show the top 5 style neighbors in CSD using our feature extractor. The green and red box around the image indicates whether or not the artist's name used to generate the SD image was present in the caption of the nearest neighbor.

Fig. 6: Top row: Images generated by Stable Diffusion. Middle and Bottom rows: Top matches retrieved by CLIP vs CSD (ours) respectively. CLIP is consistently biased towards image content, for instance retrieving image of a dog in the Column 1, 3, 4, or image of mother and baby in Column 7 or 8. Our method emphasize less on the content but more on the image styles. Please refer to the Appendix for the prompts.

Style retrieval on generated images. In Tab. 2, we show the retrieval results for *Simple* and *User-generated* prompts. We also compare our results with the second-best performing model in the previous section, CLIP ViT-L, and a recent style attribution model GDA [63]. We observe that our method outperforms CLIP on *Simple* prompt dataset. For *User-generated* prompts, the performance metrics are closer to CLIP model, but it's important to note that these prompts are inherently more complex. This complexity results in a different label distribution in the query set for the two types of prompts we examine, leading to varied metric ranges in each case. Additionally, our method consistently outperforms both baselines in content-constrained scenarios, as evidenced in Table 3. This indicates the robustness and effectiveness of our approach in dealing with a variety of prompt complexities and content specifications. We refer the reader to Appendix to understand a few caveats of this quantitative study.

Qualitative Results. In Fig. 5, we showcase a selection of Stable Diffusion-generated images alongside their top 5 corresponding matches from ContraStyles, as determined by the CSD ViT-L feature extractor. The left section of the figure displays images generated from *User-generated* prompts, while the right section includes images created from *Simple* prompts. To aid in visual analysis, matches that share a label with the query image are highlighted in green. We can clearly see that the query image and the matches share multiple stylistic elements, such as color palettes and certain artistic features like motifs or textures. We observed that in generations based on user-generated prompts, perceivable style copying typically occurs only when the prompts are shorter and contain elements that are characteristic of the artist's typical content.

In Figure 6, we present several content-constrained prompt generations and their top-1 matches based on the CLIP ViT-L/14 model versus our CSD model. We observe that the CSD model accurately matches the correct artists to queries even when there is no shared content, only style. This is evident in columns 1, 3, 4, 7, and 8, where our model,CSD matches the correct style elements despite the subjects in the images being quite different. In contrast, the CLIP model still prioritizes content, often leading to mismatches in style.

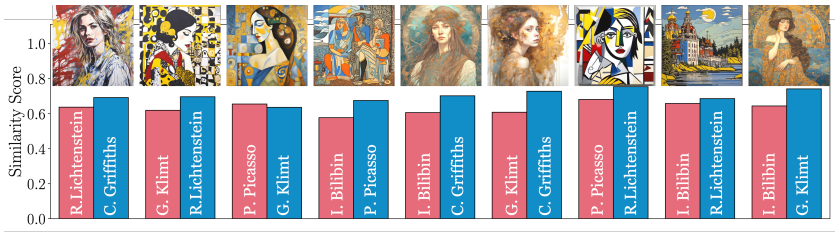

Fig. 7: Does the diffusion model prefer some styles over others? When a prompt contains two style tags, we find that SD 2.1 strongly favors the style that it can best reproduce, we suspect because of a prevalence of the style in training data. In each block, the General Style Similarity(GSS) of the left side artist (red color) is less than the right side artist (blue color). (Ref Fig. 2). The generated image is more biased towards the artist with high GSS score.

Does the model prioritize some artists over others in the prompt? So far in the study we concentrated on the impact of including an artist's name in a style transfer prompt. In this section, we present preliminary findings in scenarios where prompts feature two artists. This is inspired by real-world user prompts from Stable Diffusion and Midjourney, where prompts often include multiple artists. We used the prompt in the style `A painting in the style of <X> and <Y>`, where `X` and `Y` represent different artists. For this study, we selected five artists with varying General Style Similarity (GSS) scores (referenced in Sec. 2). The artists, ranked by descending GSS scores, are Carne Griffiths, Roy Lichtenstein, Gustav Klimt, Pablo Picasso, and Ivan Bilibin. Note that most of the chosen artists have distinct styles that significantly differ from one another.

We chose SD-XL Turbo for this analysis because it is trained on deduplicated data, reducing bias towards frequently featured artists in the train set.

The results for each pair of artists are presented in Fig. 7. Interestingly, even without specific instructions to generate a female subject, most outputs depicted women, reflecting the common subject matter of the artists studied. We also calculated the style similarity scores for each generated image, comparing them to the prototypical styles of the artists in the prompts. In most cases, the style of the artist with the higher GSS score dominated the generated image. To test for potential bias towards the artist positioned first in the prompt (`X`), we conducted two trials with reversed artist positions. The results were generally consistent, with the dominant style remaining unchanged. However, in the case of Pablo Picasso and Gustav Klimt, this pattern did not hold; the model favored Picasso's cubist style over Klimt's nouveau style, possibly due to the small difference in their GSS scores. While this is not an extensive study, a consistent trend emerged: styles of artists with higher GSS scores, like Leonid Afremov and Carne Griffiths, predominantly influenced the combined style. We leave the comprehensive study on this topic to future work.

8 Conclusion

This study proposes a framework for learning style descriptors from both labeled and unlabeled data. After building a bespoke dataset, ContraStyles, we train a model that achieves state-of-the-art performance on a range of style matching tasks, including DomainNet, WikiArt, and ContraStyles. Then, we show the substantive practical utility of this model through an investigative analysis on the extent of style copying in popular text-to-image generative models. Here, we show the model is capable of determining the factors that contribute to the frequency of style replication. Through cross-referencing of images with style copies and their original prompts, we have discover that the degree of style copying is increasing with prompt complexity. Such complex prompts lead to greater style copying compared to simple one-line prompts. This finding sheds light on the interplay between textual prompts and style transfer, suggesting that prompt design can influence the level of style copying in generative models. Finally, note that the definition of style used in this work is strictly based on artist attribution. We chose this definition because it can be operationalized and used in dataset construction. This definition is certainly not a golden truth, and we look forward to future studies using alternative, or extended definitions.

Acknowledgements

This work was made possible by the ONR MURI program and the AFOSR MURI program. Commercial support was provided by Capital One Bank, the Amazon Research Award program, and Open Philanthropy. Further support was provided by the National Science Foundation (IIS-2212182), and by the NSF TRAILS Institute (2229885).

References

1. Adobe: Firefly (2023), https://www.adobe.com/sensei/generative-ai/firefly.html
2. Agarwal, S., Karnick, H., Pant, N., Patel, U.: Genre and style based painting classification. In: 2015 IEEE Winter Conference on Applications of Computer Vision. pp. 588–594. IEEE (2015)
3. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307 (2016)
4. Bai, Z., Nakashima, Y., Garcia, N.: Explain me the painting: Multi-topic knowledgeable art description generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5422–5432 (2021)
5. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1728–1738 (2021)
6. Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., et al.: A cookbook of self-supervised learning. arXiv preprint arXiv:2304.12210 (2023)
7. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision (ICCV) (2021)
9. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 539–546. IEEE (2005)
10. Chu, W.T., Wu, Y.L.: Image style classification based on learnt deep correlation features. IEEE Transactions on Multimedia **20**(9), 2491–2502 (2018)
11. Decrypt: Greg rutkowski removed from stable diffusion but brought back by ai artists (March 2024), https://decrypt.co/150575/greg-rutkowski-removed-from-stable-diffusion-but-brought-back-by-ai-artists
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. arXiv preprint arXiv:1610.07629 (2016)
14. Gairola, S., Shah, R., Narayanan, P.: Unsupervised image style embeddings for retrieval and recognition tasks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3281–3289 (2020)
15. Garcia, N., Vogiatzis, G.: How to read paintings: semantic art understanding with multi-modal retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
16. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)
17. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)

18. Geiping, J., Goldblum, M., Somepalli, G., Shwartz-Ziv, R., Goldstein, T., Wilson, A.G.: How much data are augmentations worth? an investigation into scaling laws, invariance, and implicit regularization. arXiv preprint arXiv:2210.06441 (2022)
19. Gibson, J.J.: The senses considered as perceptual systems. (1966)
20. Graham, D.J., Hughes, J.M., Leder, H., Rockmore, D.N.: Statistics, vision, and the analysis of artistic style. Wiley Interdisciplinary Reviews: Computational Statistics **4**(2), 115–123 (2012)
21. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
22. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
23. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV). pp. 172–189 (2018)
24. Hughes, J.M., Graham, D.J., Jacobsen, C.R., Rockmore, D.N.: Comparing higher-order spatial statistics and perceptual judgements in the stylometric analysis of art. In: 2011 19th European Signal Processing Conference. pp. 1244–1248. IEEE (2011)
25. Hughes, J.M., Graham, D.J., Rockmore, D.N.: Stylometrics of artwork: uses and limitations. In: Computer Vision and Image Analysis of Art. vol. 7531, pp. 91–105. SPIE (2010)
26. Jiang, Q.Y., He, Y., Li, G., Lin, J., Li, L., Li, W.J.: Svd: A large-scale short video dataset for near-duplicate video retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5281–5289 (2019)
27. Joshi, A., Agrawal, A., Nair, S.: Art style classification with self-trained ensemble of autoencoding transformations. arXiv preprint arXiv:2012.03377 (2020)
28. Karayev, S., Trentacoste, M., Han, H., Agarwala, A., Darrell, T., Hertzmann, A., Winnemoeller, H.: Recognizing image style. arXiv preprint arXiv:1311.3715 (2013)
29. Lawrence-Lightfoot, S., Davis, J.H.: The art and science of portraiture. John Wiley & Sons (2002)
30. Lecoutre, A., Negrevergne, B., Yger, F.: Recognizing art style automatically in painting with deep learning. In: Asian conference on machine learning. pp. 327–342. PMLR (2017)
31. Lee, S., Kim, D., Han, B.: Cosmo: Content-style modulation for image retrieval with text feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 802–812 (June 2021)
32. Li, J., Yao, L., Hendriks, E., Wang, J.Z.: Rhythmic brushstrokes distinguish van gogh from his contemporaries: findings via automated brushstroke extraction. IEEE transactions on pattern analysis and machine intelligence **34**(6), 1159–1176 (2011)
33. Lu, X., Lin, Z., Shen, X., Mech, R., Wang, J.Z.: Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 990–998 (2015)
34. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4990–4998 (2017)
35. Lun, Z., Kalogerakis, E., Sheffer, A.: Elements of style: learning perceptual shape style similarity. ACM Transactions on graphics (TOG) **34**(4), 1–14 (2015)

36. Matsuo, S., Yanai, K.: Cnn-based style vector for style image retrieval. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. pp. 309–312 (2016)
37. Matthews, R.A., Merriam, T.V.: Distinguishing literary styles using neural networks. In: Handbook of neural computation, pp. G8–1. CRC Press (2020)
38. Menis-Mastromichalakis, O., Sofou, N., Stamou, G.: Deep ensemble art style recognition. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2020)
39. Midjourney: Midjourney (nd), `https://www.midjourney.com/home`
40. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems **26** (2013)
41. Park, T., Zhu, J.Y., Wang, O., Lu, J., Shechtman, E., Efros, A., Zhang, R.: Swapping autoencoder for deep image manipulation. Advances in Neural Information Processing Systems **33**, 7198–7211 (2020)
42. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1406–1415 (2019)
43. Pernias, P., Rampas, D., Richter, M.L., Pal, C., Aubreville, M.: Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In: The Twelfth International Conference on Learning Representations (2023)
44. pharmapsychotic: Clip interrogator. `https://github.com/pharmapsychotic/clip-interrogator` (2023)
45. Pizzi, E., Roy, S.D., Ravindra, S.N., Goyal, P., Douze, M.: A self-supervised descriptor for image copy detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14532–14542 (2022)
46. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
47. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
48. Rodriguez, C.S., Lech, M., Pirogova, E.: Classification of style in fine-art paintings using transfer learning and weighted image patches. In: 2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS). pp. 1–7. IEEE (2018)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
50. Ruta, D., Motiian, S., Faieta, B., Lin, Z., Jin, H., Filipkowski, A., Gilbert, A., Collomosse, J.: Aladin: all layer adaptive instance normalization for fine-grained style similarity. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11926–11935 (2021)
51. Sablatnig, R., Kammerer, P., Zolda, E.: Hierarchical classification of paintings using face-and brush stroke models. In: Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170). vol. 1, pp. 172–174. IEEE (1998)
52. Saleh, B., Elgammal, A.: Large-scale classification of fine-art paintings: Learning the right metric on the right feature. arXiv preprint arXiv:1505.00855 (2015)

53. Sandoval, C., Pirogova, E., Lech, M.: Two-stage deep learning approach to the classification of fine-art paintings. IEEE Access **7**, 41770–41781 (2019)
54. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. arXiv preprint arXiv:2210.08402 (2022)
55. Silva, J.M., Pratas, D., Antunes, R., Matos, S., Pinho, A.J.: Automatic analysis of artistic paintings using information-based measures. Pattern Recognition **114**, 107864 (2021)
56. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
57. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: 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)
58. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Understanding and mitigating copying in diffusion models. Advances in Neural Information Processing Systems **36**, 47783–47803 (2023)
59. Srinivasa Desikan, B., Shimao, H., Miton, H.: Wikiartvectors: style and color representations of artworks for cultural analysis via information theoretic measures. Entropy **24**(9), 1175 (2022)
60. Tenenbaum, J., Freeman, W.: Separating style and content. Advances in neural information processing systems **9** (1996)
61. Walmer, M., Suri, S., Gupta, K., Shrivastava, A.: Teaching matters: Investigating the role of supervision in vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
62. Wang, J., Yang, H., Fu, J., Yamasaki, T., Guo, B.: Fine-grained image style transfer with visual transformers. In: Proceedings of the Asian Conference on Computer Vision. pp. 841–857 (2022)
63. Wang, S.Y., Efros, A.A., Zhu, J.Y., Zhang, R.: Evaluating data attribution for text-to-image models. arXiv preprint arXiv:2306.09345 (2023)
64. Wilber, M.J., Fang, C., Jin, H., Hertzmann, A., Collomosse, J., Belongie, S.: Bam! the behance artistic media dataset for recognition beyond photography. In: Proceedings of the IEEE international conference on computer vision. pp. 1202–1211 (2017)
65. Willats, J., Durand, F.: Defining pictorial style: Lessons from linguistics and computer graphics. Axiomathes **15**(3), 319–351 (2005)
66. Wynen, D., Schmid, C., Mairal, J.: Unsupervised learning of artistic styles with archetypal style analysis. Advances in Neural Information Processing Systems **31** (2018)
67. Yao, L., Li, J., Wang, J.Z.: Characterizing elegance of curves computationally for distinguishing morrisseau paintings and the imitations. In: 2009 16th IEEE International Conference on Image Processing (ICIP). pp. 73–76. IEEE (2009)
68. Zhang, W., Cao, C., Chen, S., Liu, J., Tang, X.: Style transfer via image component analysis. IEEE Transactions on multimedia **15**(7), 1594–1601 (2013)