A Unified Arbitrary Style Transfer Framework via Adaptive Contrastive Learning

YUXIN ZHANG, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China
FAN TANG, Institute Of Computing Technology, CAS, China
WEIMING DONG, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China
HAIBIN HUANG and CHONGYANG MA, Kuaishou Technology, China
TONG-YEE LEE, Department of Computer Science and Information Engineering, National Cheng-Kung University, Taiwan
CHANGSHENG XU, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China

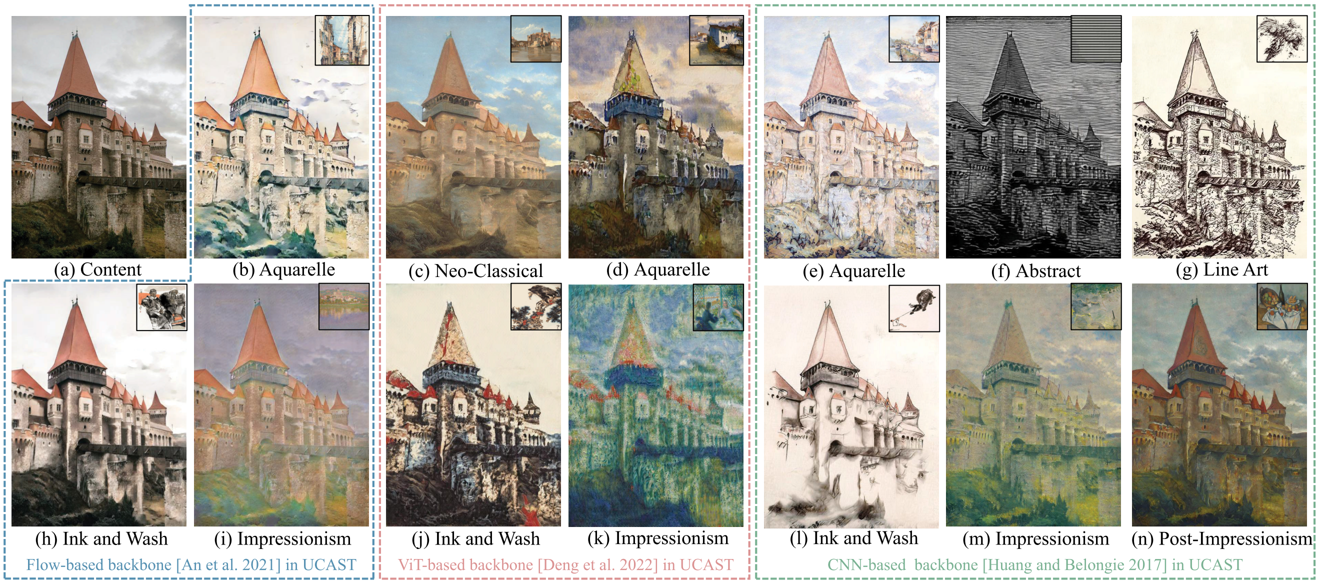


Fig. 1. Style transfer results of three different generative backbones trained under our framework, which can robustly and effectively handle various painting styles. The input content image is shown in (a). The style reference is shown as the inset for each result. Our method can faithfully capture the style of each painting and generate a result with a unique artistic visual appearance. Content image credit: Julia Volk/Pexels (Free to use) [Pexels 2023]. Style image credits: (c) Jean-Baptiste-Camille Corot/National Gallery of Art (CC0), {(i) Claude Monet, (k)Pierre-Auguste Renoir, (m) Utagawa Hiroshige, (n) Paul Cezanne}/The Art Institute of Chicago (CC0) [Art Institute of Chicago 2023].

This work presents Unified Contrastive Arbitrary Style Transfer (UCAST), a novel style representation learning and transfer framework, that can fit in most existing arbitrary image style transfer models, such as CNN-based, ViT-based, and flow-based methods. As the key component in image style transfer tasks, a suitable style representation is essential to achieve satisfactory results. Existing approaches based on deep neural networks typically use second-order statistics to generate the output. However, these hand-crafted features computed from a single image cannot leverage style information sufficiently, which leads to artifacts such as local distortions and style inconsistency. To address these issues, we learn style representation directly from a large number of images based on contrastive

This work was supported in part by National Key R&D Program of China under no. 2020AAA0106200, by National Natural Science Foundation of China under nos. 61832016, 62102162, and U20B2070, and in part by Beijing Natural Science Foundation under no. L221013, in part by the National Science and Technology Council, under no. 111-2221-E-006-112-MY3, Taiwan.

Authors' addresses: Y. Zhang, W. Dong, and C. Xu, MAIS, Institute of Automation, Chinese Academy of Sciences and School of Artificial Intelligence, University of Chinese Academy of Sciences, 95 Zhongguancun East Road, Haidian District, Beijing, China, 100190; emails: {zhangyuxin2020, weiming.dong}@ia.ac.cn, csxu@nlpr.ia.ac.cn; F. Tang, Institute of Computing Technology, Chinese Academy of Sciences, No. 6 Kexueyuan South Road Zhongguancun, Haidian District, Beijing, China, 100190; email: tangfan@ict.ac.cn; H. Huang and C. Ma, Kuaishou Technology, No. 6 Shangdi West Road, Haidian District, Beijing, China, 100085; emails: {huanghaibin03, chongyangma}@kuaishou.com; T.-Y. Lee, Department of Computer Science and Information Engineering, National Cheng-Kung University, No. 1, University Road, Tainan City 701, Taiwan, 701401; email: tonylee@ncku.edu.tw.

0730-0301/2023/07-ART169 $15.00
https://doi.org/10.1145/3605548

learning by considering the relationships between specific styles and the holistic style distribution. Specifically, we present an adaptive contrastive learning scheme for style transfer by introducing an input-dependent temperature. Our framework consists of three key components: a parallel contrastive learning scheme for style representation and transfer, a domain enhancement (DE) module for effective learning of style distribution, and a generative network for style transfer. Qualitative and quantitative evaluations show the results of our approach are superior to those obtained via state-of-the-art methods. The code is available at https://github.com/zyxElsa/CAST_pytorch.

CCS Concepts: • **Computing methodologies → Image processing;**

Additional Key Words and Phrases: Arbitrary style transfer, contrastive learning, style encoding

ACM Reference format:
Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. 2023. A Unified Arbitrary Style Transfer Framework via Adaptive Contrastive Learning. *ACM Trans. Graph.* 42, 5, Article 169 (July 2023), 16 pages.
https://doi.org/10.1145/3605548

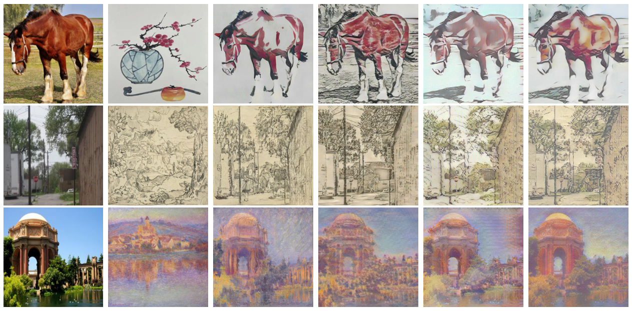


Fig. 2. Comparison with the latest style transfer methods: CNN-based method AdaAttN [Liu et al. 2021b], neural flow-based method ArtFlow [An et al. 2021], and ViT-based method StyTr^2 [Deng et al. 2022], all of which rely on second-order statistics. Our method can faithfully transfer styles while ensuring structural consistency with the content images. Content image credit (the 1st row): Pixabay/Pexels (CC0) [Pexels 2023]. Style image (the 2nd and 3rd rows) credit: {Michel Ange Corneille, Claude Monet}/AIC (CC0) [Art Institute of Chicago 2023].

1 INTRODUCTION

If a picture is worth a thousand words, then an artwork may tell the whole story. The art style depicts the visual appearance of an artwork and characterizes how the artist expresses a theme and shows his/her creativity. The features that identify an artwork, such as the artist's use of strokes, color, and composition, determine the style [McArdle 2022]. Artistic style transfer, as an efficient way to create a new painting by combining the content of natural images and the style of an existing painting image, is a major research topic in computer graphics and computer vision [Jing et al. 2020b].

The main challenges of arbitrary style transfer are extracting styles from artistic images and mapping a specific realistic image into an artistic one in a controllable way. The core problem for style extraction is to find an effective representation of styles because providing explicit definitions across different styles is difficult in general. To build a reasonable style feature space, exploring the relationship and distribution of styles is necessary to capture individual and holistic characteristics. For the mapping, several generative mechanisms are adopted to address different issues, such as autoencoder [Huang and Belongie 2017; Liu et al. 2021b], neural flow model [An et al. 2021], and visual transformer [Deng et al. 2022]. In contrast to the goal of those methods, this article proposes to improve arbitrary style transfer via a unified framework that offers the guidance of proper artistic style representation and works for various generative backbones.

Since Gatys et al. [2016] proposed to use the Gram matrix as an artistic style representation, high-quality visual results are generated by advanced neural style transfer networks. Despite remarkable progress in the field of arbitrary image style transfer, the second-order feature statistics (Gram matrix or mean/variance) style representation has restricted the further development and application. In Figure 1, the appearances of different artwork styles vary considerably in terms of not only the colors and local textures but also the layouts and compositions. Figure 2(d)–(f) shows the results of three recently proposed state-of-the-art style transfer approaches. Aligning the distributions of neural activation between images using second-order statistics results in difficulty in capturing the color distribution or the spatial layouts or imitating the specific detailed brush effects of different styles.

In this article, the core problem for neural style transfer, that is, the proper artistic style representation, is revisited. The widely used second-order statistics as a global style descriptor can distinguish styles to some extent, but they are not the optimal way to represent styles. By second-order statistics, arbitrary stylization formulates styles through artificially designed image features and loss functions in a heuristic manner. The network learns to fit the second-order statistics of the style image and the generated image, instead of the style itself. Our key insight is that a person without artistic knowledge has difficulty defining the style if only one artistic image is given, but identifying the difference between dissimilar styles is relatively easy. Therefore, exploring the relationship and distribution of styles directly from artistic images instead of using pre-defined style representations is worthwhile. This article proposes to improve arbitrary style transfer with a novel style representation based on contrastive learning. Specifically, this work presents a **Unified Contrastive Arbitrary Style Transfer** (**UCAST**) framework for image style representation learning and style transfer, which consists of a generative backbone, a parallel contrastive learning scheme, and a **domain enhancement** (**DE**) module. Contrastive learning is introduced to consider the positive and negative relationships between different styles, and DE is used to learn the overall domain distribution of artistic images. UCAST can be plug-and-play for most arbitrary image style transfer methods to improve their performance.

Given that different images may share similar styles, considering similar styles is necessary in the style modeling, and the style contrastive learning should tolerate highly similar samples. Moreover, compared with per-style-per-model methods and multiple-style-per-model tasks, arbitrary image style transfer has the difficulty that when dealing with specific content-style pairs, the content image and style image may not always be compatible with each other. For instance, when using a realistic image with a large smooth area as the content and an artistic image with rich texture as the style, undesirable artifacts may be observed in the stylization output. Thus, an adaptive contrastive loss that is implemented with a novel dual input-dependent temperature scheme is proposed. Our adaptive contrastive loss considers the similarities between the target style image and other artistic images as well as the similarities between the target style image and the input content image to address the above problems.

This work extends the conference paper "Domain Enhanced Arbitrary Image Style Transfer via Contrastive Learning" which is published in ACM SIGGRAPH 2022 [Zhang et al. 2022b]. The style transfer framework is improved with a novel parallel adaptive contrastive learning scheme with two temperature values instead of one previously. The conference version is extended with comprehensive experiments and demonstrates that our unified framework UCAST can significantly improve the quality of the stylized results for existing arbitrary image style transfer models. Furthermore, our method is applied to video-style transfer.

Our contribution can be summarized as follows:

- A novel framework called UCAST, which can easily integrate various types of style transfer backbones and lead to improved visual quality in the stylization results, is proposed.
- A novel style representation learning method via contrastive learning without employing the commonly used second-order statistics of image features is proposed. Contrastive learning and DE are introduced by considering the relationships between styles as well as the global distribution of styles, which solves the problem that existing style transfer models cannot effectively leverage a large amount of available artistic images.
- Adaptive contrastive learning for arbitrary style transfer tasks, which allows the model to be tolerant to similar styles and improve the robustness of various content-style inputs, is proposed.

2 RELATED WORK

Image style transfer. Traditional style transfer methods such as stroke-based rendering [Fišer et al. 2016] and image filtering [Wang et al. 2004] typically use low-level hand-crafted features. Gatys et al. [2016] and the follow-up variants [Gatys et al. 2017; Kolkin et al. 2019] demonstrate that the statistical distribution of features extracted from pre-trained deep convolutional neural networks can effectively capture style patterns. Although the results are remarkable, these methods formulate the task as a complex optimization problem, which leads to high computational cost. Some recent approaches rely on a learnable neural network to match the statistical information in feature space for efficiency. Per-style-per-model methods [Johnson et al. 2016; Kwon and Ye 2022; Puy and Pérez 2019; Ulyanov et al. 2016; Zhang et al. 2023b] train a specific network for each individual style. Multiple-style-per-model methods [Chen et al. 2017; Dumoulin et al. 2017; Gao et al. 2020; Zhang and Dana 2018] represent multiple styles using a single model.

Arbitrary style transfer methods [Deng et al. 2022, 2020; Li et al. 2017; Liao et al. 2017; Svoboda et al. 2020; Wu et al. 2021a; Zhang et al. 2022b] build more flexible feed-forward architectures to handle an arbitrary style using a unified model. AdaIN [Huang and Belongie 2017] and DIN [Jing et al. 2020a] directly align the overall statistics of content features with the statistics of style features and adopt conditional instance normalization. However, the dynamic generation of affine parameters in the instance normalization layer may cause distortion artifacts. Instead, several methods follow the encoder-decoder manner, where feature transformation and/or fusion is introduced into an autoencoder-based framework. For instance, Lee et al. [2018] propose to embed images onto two spaces and present an approach based on disentangled representation for producing diverse outputs without paired training images. Li et al. [2019] achieve universal style transfer by developing a cross-domain feature **linear transformation matrix (LST)** and decoding from the transformed features. Park et al. [2019] provide a flexible mapping of the semantically nearest style features onto the content features by SANet. Park et al. [2020] propose the Swapping Autoencoder that can encode an image into two independent components and enforce that any swapped combination maps to a realistic image. Deng et al. [2021] propose MCCNet for efficient video style transfer by fusing input content features and style features via multichannel correlation. Liu et al. [2021b] present an **adaptive attention normalization (AdaAttN)** module to consider both shallow and deep features for attention score calculation. Wang et al. [2022] propose an **aesthetic-enhanced universal style transfer (AesUST)** approach that incorporates the aesthetic features to enhance the style transfer process and can generate aesthetically more realistic and pleasing results. GAN-based methods [Kotovenko et al. 2019a, b; Sanakoyeu et al. 2018a; Svoboda et al. 2020; Zhu et al. 2017] have been successfully used in collection style transfer, which considers style images in a collection as a domain [Chen et al. 2021b; Lin et al. 2021; Wang et al. 2023; Xu et al. 2021]. An et al. [2021] propose reversible neural flows and an unbiased feature transfer module (ArtFlow) to prevent content leaks during universal style transfer. Inspired by the breakthrough of **visual transformer (ViT)**, many researchers have developed ViT for style transfer tasks. Wu et al. [2021a] propose a feed-forward style transfer method (StyleFormer) that includes a transformer-driven style composition module. Deng et al. [2022] propose a ViT-based style transfer method (StyTr^2) that considers the long-range dependencies of input images to avoid biased content representation. Zhang et al. [2022a] performed exact matching of feature distributions and apply this method to arbitrary style transfer.s Benefitting from the pre-trained text-to-image generative models [2022], researchers have adopted diffusion models for style transfer tasks [Huang et al. 2022a, b; Zhang et al. 2023a]. Zhang et al. [2023a] propose an **inversion-based style transfer (InST)** method , which can efficiently and accurately learn the key information of an image, thus capturing and transferring the artistic style of a painting without providing complex textual descriptions.

Contrastive learning. Contrastive learning has been used in many applications, such as image dehazing [Wu et al. 2021b], context prediction [Santa Cruz et al. 2019], geometric prediction [Liu et al. 2019], and image translation. Contrastive learning is introduced in image translation to preserve the content of the input [Han et al. 2021] and reduce mode collapse [Jeong and Shin 2021; Kang and Park 2020; Liu et al. 2021a]. CUT [Park et al. 2020a] proposes patch-wise contrastive learning by cropping

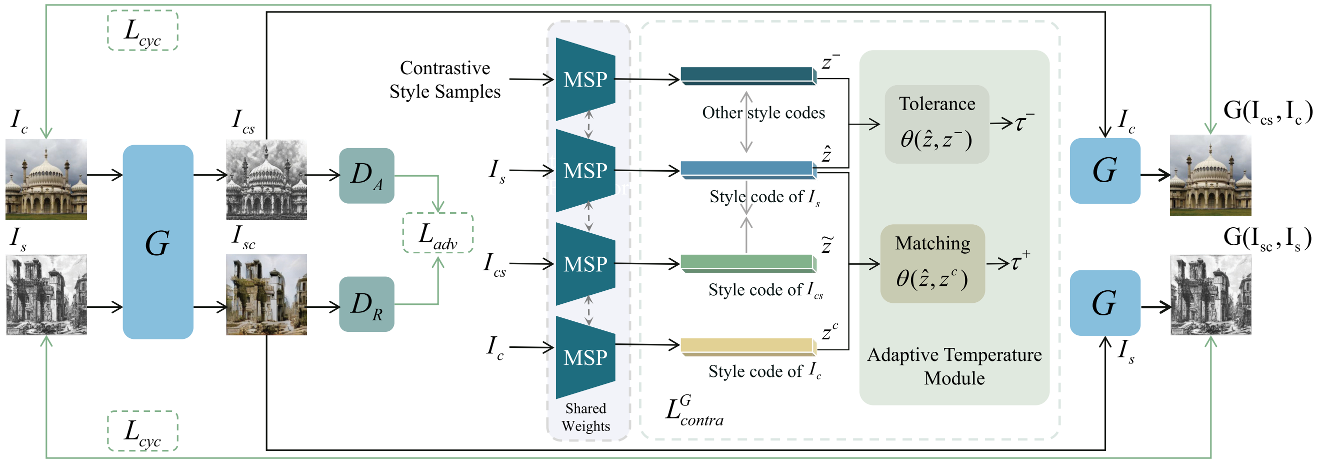


Fig. 3. UCAST consists of a generator G, a parallel contrastive learning scheme relying on a MSP module, and a DE module. The generator is given the content image I_c and the style image I_s and generates images I_{cs} and I_{sc}. Then, I_{cs} and I_s are fed into the MSP module to generate the corresponding style code $\tilde{\mathbf{z}}$ and $\hat{\mathbf{z}}$, which are used as positive samples in the style contrastive learning process. The style codes $\mathbf{z}^-$ of other artistic images in the style bank are used as negative samples. I_c is fed into the MSP module to generate the corresponding style code $\mathbf{z}^c$. We design an adaptive temperature module that computes the temperature τ^+ of the positive sample and the temperature τ^- of the negative samples on the style codes. The contrastive style loss $\mathcal{L}^G_{contra}$ is computed on the temperatures and the style codes. The DE module is based on the adversarial loss $\mathcal{L}_{adv}$ and the cycle consistency loss $\mathcal{L}_{cyc}$. Style image credit: Giovanni Battista Piranesi/AIC (CC0) [Art Institute of Chicago 2023].

input and output images into patches and maximizing the mutual information between patches. Following CUT, TUNIT [Baek et al. 2021] adopts contrastive learning on images with similar semantic structures. However, the semantic similarity assumption does not hold for arbitrary style transfer tasks, which leads the learned style representations to a significant performance drop. IEST [Chen et al. 2021a] applies contrastive learning to image style transfer based on feature statistics (mean and standard deviation) as style priors. The contrastive loss is calculated only within the generated results. Contrastive learning in IEST is an auxiliary method to associate stylized images sharing the same style, and the ability comes from the feature statistics from pre-trained VGG. CCPL [Wu et al. 2022] introduces contrastive learning for video style transfer by considering the frame-wise patch differences. Differently, contrastive learning for style representation is introduced here by proposing a novel framework that uses visual features comprehensively to represent style for the task of arbitrary image style transfer.

Temperature is a critical parameter for the success of a contrastive-learning-based method. Wang and Liu [2021] show that the contrastive loss has a hardness-aware property, which makes contrastive learning naturally focus on difficult negative samples. Such hardness awareness helps learn separable, uniformly distributed features but also leads to the low tolerance of semantically similar samples. The extent of penalties on hard negative samples is determined by temperature τ. As the temperature decreases, the relative penalty concentrates more on the high-similarity region, whereas as the temperature increases, the relative penalty distribution becomes more uniform, which means all negative samples are penalized equally. Relations are built between uniformity, tolerance, and temperature. Zhang et al. [2022b] introduced vector decomposition for analyzing the collapse issue based on gradient analysis of the l_2-normalized representation vector and proposed a unified perspective on how negative samples and simple Siamese method alleviate collapse. Caron et al. [2021] investigated dual temperature from the perspective of knowledge distillation and proposed a simple self-supervised method, in which the teacher adopts a lower temperature than the student to help in knowledge distillation. Zhang et al. [2021] learn temperature as an input-dependent variable. They consider temperature as a measure of embedding confidence and propose temperature as uncertainty. Zhang et al. [2022a] adopt dual temperature in a contrastive InfoNCE for realizing independent control of two hardness-aware sensitiveness. Previous temperature analysis works mainly focus on the penalty's unevenness of negative samples within an anchor or the sum of penalties of different anchors within a training batch. By contrast, this work simultaneously considers the proportion of penalties between the positive sample and negative samples.

3 METHOD

3.1 Overview

Our unified framework for arbitrary image style transfer as a separated network structure can be plug-and-play for most arbitrary image style transfer models. As shown in Figure 3, our UCAST consists of three key components: (1) a parallel contrastive learning scheme that is applied to the style representation learning and the style transfer process; (2) a DE scheme to further help learn the distribution of the artistic image domain and (3) a generator G to generate the stylization output. (1) and (2) are used for learning style features to measure the difference between artistic images and realistic images. The parallel contrastive learning scheme focuses on forcing the specific reference artistic image and the generated result to have the same style, whereas the DE scheme pays attention to the holistic difference between the artistic domain and the realistic domain.

The main structure of our parallel contrastive learning scheme is a **multilayer style projector** (**MSP**) trained to project features of artistic images into style codes. The contrastive losses are introduced to guide parallel optimization processes, including the training of MSP and the generator. When training the generator, adaptive contrastive loss implemented with dual input-dependent temperature is introduced. By considering the similarities between the style codes of the reference style image and other artistic images, our adaptive contrastive loss is more tolerant to style-consistent samples. The input-dependent temperature is also influenced by the similarities between the style codes of the target style image and the input content image, to increase the robustness of various content-style pairs and prevent artifacts. The DE scheme is accomplished by two discriminators for the artistic domain and the realistic domain. Adversarial loss helps the discriminator model the distribution of the corresponding domain, and cycle consistency loss is adopted to maintain the content information.

3.2 Parallel Contrastive Learning

3.2.1 Multilayer Style Projector. Our goal is to develop a unified arbitrary style transfer framework that can capture and transfer the local stroke characteristics and overall appearance of an artistic image to a natural image. A key component is to find a suitable style representation that can be used to distinguish different styles and further guide the generation of style images. To this end, an MSP module, which includes a style feature extractor and a multilayer projector, is designed. Instead of using features from a specific layer or a fusion of multiple layers, our MSP projects features of different layers into separate latent style spaces to encode local and global style cues.

Specifically, VGG-19 [Simonyan and Zisserman 2014] is adopted, and the VGG-19 model pre-trained on ImageNet with a collection of 18,000 artistic images in fifty categories is finetuned. M layers of feature maps in VGG-19 are selected as input to our multilayer projector (layers of ReLU1_2, ReLU2_2, ReLU3_3, and ReLU4_3 are used in all experiments). Max pooling and average pooling are used to capture the mean and peak values of features. The multilayer projector consists of pooling, convolution, and several multilayer perceptron layers, and it projects the style features into a set of K-dimensional latent style code, as shown in Figure 4.

After training, MSP can encode an artistic image into a set of latent style code $\{\mathbf{z}_i | i \in [1, M], \mathbf{z}_i \in \mathbb{R}^K\}$, which can be plugged into an existing style transfer network (i.e., replacing the mean and variance in AdaIN [Huang and Belongie 2017]) as the guidance for stylization. Next, how to jointly train MSP and style transfer networks with a contrastive learning strategy is described.

3.2.2 Contrastive Style Representation Learning. A branch of the parallel contrastive learning scheme is style representation learning. The MSP needs to be trained to obtain a reasonable style representation that is in the form of the style code $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M\}$. However, the ground-truth style code for supervised training is lacking. Therefore, contrastive learning is adopted, and a new contrastive style loss is designed as an implicit measurement for the MSP training.

When training the MSP module, an image I and its augmented version I^+ (random resizing, cropping, and rotations) are fed into an M-layer style feature extractor, which is the pre-trained VGG-19 network. The extracted style features are then sent to the multilayer projector, which is an M-layer neural network and maps the style features to a set of K-dimensional vectors $\{\mathbf{z}\}$. The contrastive representation learns the visual styles of images by maximizing the mutual information between I and I^+ in contrast to other artistic images within the dataset considered as negative samples $\{I^-\}$. Specifically, the images I, I^+, and N negative samples are mapped into M groups of K-dimensional vectors $\mathbf{z}, \mathbf{z}^+ \in \mathbb{R}^K$ and $\{\mathbf{z}^- \in \mathbb{R}^K\}$. The vectors are normalized to prevent collapsing, respectively. A large dictionary of 4,096 negative examples is maintained using a memory bank architecture following MOCO [He et al. 2020]. The negative examples are sampled from the memory bank. Following [Van den Oord et al. 2019], the contrastive loss function is defined to train our MSP module as follows:

$$\mathcal{L}^{MSP}_{contra} = -\sum_{i=1}^{M} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_i^+/\tau)}{\exp(\mathbf{z}_i \cdot \mathbf{z}_i^+/\tau) + \sum_{j=1}^{N} \exp(\mathbf{z}_i \cdot \mathbf{z}_{i_j}^-/\tau)}, \quad (1)$$

where $\cdot$ denotes the dot product of two vectors. The contrastive loss between *images* is calculated, as opposed to CUT [Park et al. 2020a] that adopts contrastive learning by cropping images into patches and maximizing the mutual information between *patches*.

3.2.3 Contrastive Style Transfer. The other branch of the parallel contrastive learning scheme is the style transfer process. The above contrastive representation provides a proper measurement for the generator G to transfer styles between images. The loss is computed using the contrastive representations of the output image I_{cs} and the reference style image I_s, then I_{cs} has a style similar to I_s:

$$\mathcal{L}^{G}_{contra} = -\sum_{i=1}^{M} \log \frac{\exp(\tilde{\mathbf{z}}_i \cdot \hat{\mathbf{z}}_i/\tau)}{\exp(\tilde{\mathbf{z}}_i \cdot \hat{\mathbf{z}}_i/\tau) + \sum_{j=1}^{N} \exp(\tilde{\mathbf{z}}_i \cdot \mathbf{z}_{\mathbf{i_j}}^-/\tau)}, \quad (2)$$

where $\tilde{\mathbf{z}}$ and $\hat{\mathbf{z}}$ denote the contrastive representation of I_{cs} and I_s, respectively. The specific generated and reference images are taken as positive examples, and contrastive loss is utilized as guidance to transfer styles, which is a one-on-one process. Differently, the contrastive loss in IEST [Chen et al. 2021a] is calculated only within generated results, and it takes a set of images as positive examples, which could reduce the style consistency with the given reference (see Figure 7).

3.2.4 Adaptive Contrastive Learning. The model needs to tolerate these style similarities because different artworks could have similar styles. Contrastive learning seeks to minimize the distance between positive samples and maximize the distance between negative samples in the representation space. By gradient analysis, [Wang and Liu 2021] demonstrate the gradients with regard to negative samples are proportional to the similarity between the particular negative sample and the anchor, proving the contrastive loss is a hardness-aware loss function. Temperature τ controls the distribution of negative gradients. Smaller temperatures tend to focus more on the anchor point's nearest neighbors, whereas larger temperatures penalize negative samples equally. When the temperature is fixed, the gradient's magnitude with respect to a

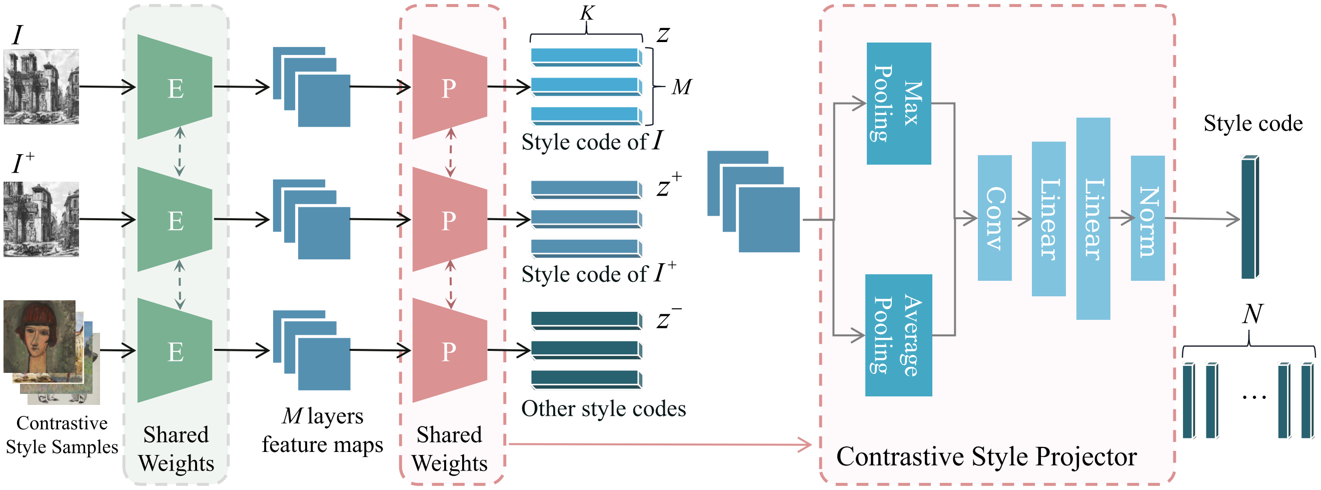
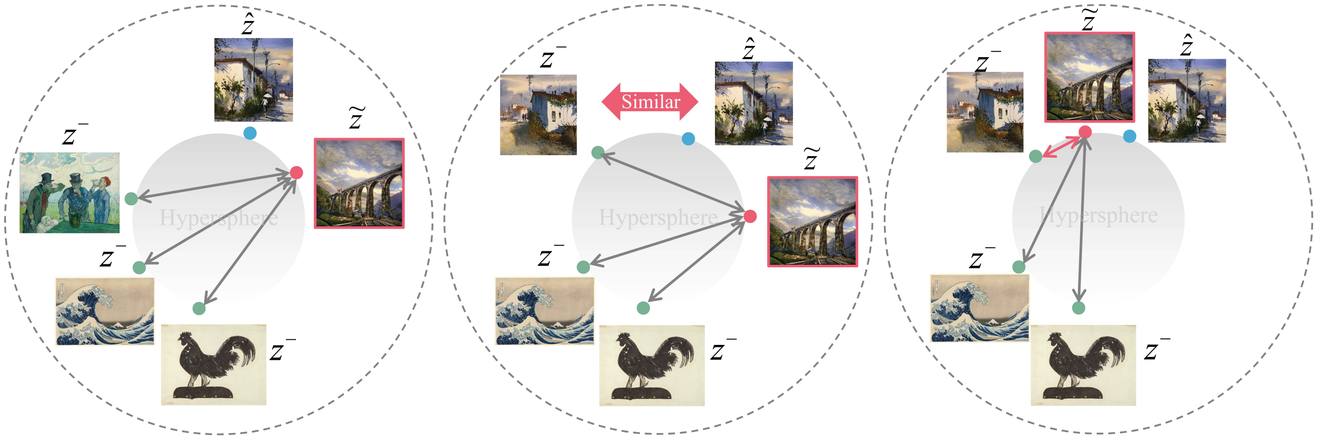


Fig. 4. Overview of our MSP module, which includes a VGG-19-based style feature extractor E and a multilayer projector P. P maps the extracted features to style codes $\{\mathbf{z}\}$ that are then saved in the style memory bank. Image credits: {Giovanni Battista Piranesi, Amedeo Modigliani}/AIC (CC0) [Art Institute of Chicago 2023].



(a) Fixed temperature: Case I

(b) Fixed temperature: Case II

(c) Adaptive temperature: Case II

Fig. 5. Visualization of the embedding distribution of artistic images and generated results on a hypersphere. Style image credits (from left to right) ({Vincent van Gogh, Katsushika Hokusai}/AIC (CC0) [Art Institute of Chicago 2023], Nicholas Acampora/NGA (CC0) [National Gallery of Art 2023].

positive sample is equal to the sum of gradients with respect to all negative samples. Prior works of temperature analysis mainly focus on the penalty's unevenness of negative samples within an anchor [Wang and Liu 2021] or the sum of penalties of different anchors within a training batch [Zhang et al. 2022a]. Differently, this work pays attention to the proportion of penalties between the positive sample and negative samples.

Figure 5 shows the embedding distribution with four real paintings and one generated image on a hypersphere. Figure 5(a) shows that when the style of the reference image and the other artistic images served as negative samples vary differently, the punishment of the fixed small temperature may work well. Different artistic images may share similar styles. When similar style images act as negative samples, as shown in Figure 5(b), the ideal embedding of the generated image is separated from all the negative samples but closer to the similar negative samples. However, the contrastive loss with a fixed small temperature provides strong punishment on similar samples due to the hardness-aware attribute, which means the generated image may be pushed away from the similar negative sample too much, which is not a reasonable embedding in the hypersphere. Our adaptive contrastive style transfer approach is aware of the negative samples that share a similar style with the reference image. When high-similarity negative samples appear, our approach will gain tolerance by increasing the temperature accordingly. Figure 5(c) shows that, with the help of our adaptive contrastive style transfer approach, the generator is guided under a reasonable loss, and the generated image can achieve a better embedding.

To further illustrate our adaptive temperature mechanism, the similarities of the positive sample and the negative samples in

Equation (2) are substituted with $\mathbf{s}_{\mathbf{i}}^{+} = \tilde{\mathbf{z}}_i \cdot \hat{\mathbf{z}}_i$, $\mathbf{s}_{\mathbf{i_j}}^{-} = \tilde{\mathbf{z}}_i \cdot \mathbf{z}_{\mathbf{i_j}}^{-}$:

$$\mathcal{L}_{contra}^{G} = -\sum_{i=1}^{M} \log \frac{\exp(\mathbf{s}_{\mathbf{i}}^{+}/\tau^{+})}{\exp(\mathbf{s}_{\mathbf{i}}^{+}/\tau^{+}) + \sum_{j=1}^{N} \exp(\mathbf{s}_{\mathbf{i_j}}^{-}/\tau^{-})}, \tag{3}$$

where τ^{+} and τ^{-} indicate the temperatures of the positive samples and the negative samples, respectively. The gradients are analyzed with respect to positive samples and different negative samples. Specifically, the gradients with respect to the positive similarity $\mathbf{s}_{\mathbf{i}}^{+}$ and the negative similarity $\mathbf{s}_{\mathbf{i_j}}^{-}$ are formulated as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}_{contra}^{G}}{\partial \mathbf{s}_{\mathbf{i}}^{+}} &= -\sum_{i=1}^{M} \frac{1}{\tau^{+}} \cdot \frac{\sum_{j=1}^{N} \exp(\mathbf{s}_{\mathbf{i_j}}^{-}/\tau^{-})}{\exp(\mathbf{s}_{\mathbf{i}}^{+}/\tau^{+}) + \sum_{j=1}^{N} \exp(\mathbf{s}_{\mathbf{i_j}}^{-}/\tau^{-})}, \\ \frac{\partial \mathcal{L}_{contra}^{G}}{\partial \mathbf{s}_{\mathbf{i_j}}^{-}} &= -\sum_{i=1}^{M} \frac{1}{\tau^{-}} \cdot \frac{\exp(\mathbf{s}_{\mathbf{i_j}}^{-}/\tau^{-})}{\exp(\mathbf{s}_{\mathbf{i}}^{+}/\tau^{+}) + \sum_{j=1}^{N} \exp(\mathbf{s}_{\mathbf{i_j}}^{-}/\tau^{-})}. \end{aligned} \tag{4}$$

Equation (4) shows the magnitude of the gradient with respect to the positive sample is proportional to the sum of gradients with respect to all the negative samples. By controlling τ^{-} and τ^{+}, the strength of penalties on the positive sample and negative samples can be changed.

This work proposes an input-dependent scheme to determine temperature by considering the similarities between the style code of the reference style $\hat{\mathbf{z}}$ and the style codes of other artistic images $\mathbf{z}_{\mathbf{i_j}}^{-}$. The more highly similar samples the memory bank contains, the larger the temperature is. To achieve this, the sigmoid function, which is a monotonic function with upper and lower bounds, is used to represent temperature. Given that the sigmoid function is centered at the point of the independent variable with a value of 0, the image similarity (the independent variable of the sigmoid function) needs to be normalized to a distribution with a mean of 0. The distribution of image similarity is assumed to follow a Gaussian distribution. The mean and variance of the image similarity during training are then calculated to normalize it. During training, the mean and variance of the distribution of the data are approximated as the number of samples increases. The recursive rules are as follows: The new mean is obtained by weighting the average similarity of each new image with the known mean similarity and then updating the average. Similarly, the new variance is derived by weighting the difference between each new image similarity and the known mean similarity with the known variance, and then updating the variance. Our input-dependent temperature is computed as follows:

$$\begin{aligned} \tau^{-} &= t_{range}^{-} \cdot \frac{1}{1 + \exp(-(\sum_{j=1}^{N} g(\mathbf{s}_{\mathbf{i_j}}^{-}) - \mu^{-}) \cdot \sigma^{-})} + t_{bound}^{-}, \\ g(\mathbf{s}_{\mathbf{i_j}}^{-}) &= \begin{cases} \mathbf{s}_{\mathbf{i_j}}^{-} & \text{for} \quad \mathbf{s}_{\mathbf{i_j}}^{-} > \mathbf{s}^{-} \\ 0 & \text{for} \quad \mathbf{s}_{\mathbf{i_j}}^{-} \leq \mathbf{s}^{-} \end{cases}, \end{aligned} \tag{5}$$

where μ^{-} and σ^{-} indicate the estimation of the mean and standard deviation of $\sum_{j=1}^{N} g(\mathbf{s}_{\mathbf{i_j}}^{-})$, respectively. t_{range}^{-} and t_{bound}^{-} denote the range and lower bound of τ^{-}. The commonly used temperature variation in contrastive learning is used. t_{range}^{-} is set to 1 and t_{bound}^{-} is set to 0.05.

Arbitrary style transfer task often has the problem that the style images may not always be suitable for the content image and thus, increase undesired artifacts. For example, when transferring a texture-rich style to a smooth content image, the model may produce artifacts and distortion (e.g., the 4th row of Figure 7). Therefore, various content-style pairs must be adaptively handled to increase the robustness. To overcome the said problem, a suitability-aware scheme is proposed to determine the temperature based on the similarity between the style code of the reference image $\hat{\mathbf{z}}$ and the style code of the content image $\mathbf{z}_i^c$. When the reference style and the content image are dissimilar, the penalty is assigned more to negative samples to prevent artifacts from being overly stylized:

$$\begin{aligned} \tau^{+} &= \tau^{-} \cdot f(\hat{\mathbf{z}}, \mathbf{z}_i^c), \\ f(\hat{\mathbf{z}}, \mathbf{z}_i^c) &= t_{range}^{+} \cdot \frac{1}{1 + \exp((\hat{\mathbf{z}} \cdot \mathbf{z}_i^c - \mu^{+}) \cdot \sigma^{+})} + t_{bound}^{+}, \end{aligned} \tag{6}$$

where μ^{+} and σ^{+} indicate the estimation of the mean and standard deviation of $\hat{\mathbf{z}} \cdot \mathbf{z}_i^c$), respectively. $t_{range}^{+} = 1$ and $t_{bound}^{+} = 0.5$ denote the range and lower bound of the scale factor of τ^{+}.

3.3 Domain Enhancement

DE with adversarial loss is introduced to enable the network to learn the style distribution. Recent style transfer models employ GAN [Goodfellow et al. 2014] to align the distribution of generated images with specific artistic images [Chen et al. 2021b; Lin et al. 2021]. The adversarial loss can enhance the holistic style of the stylization results while it strongly relies on the distribution of datasets. Even with the specific artistic style loss, the generation is often not robust enough to be artifact-free.

Differently from these previous methods, the images in the training set are divided into a realistic domain and an artistic domain, and two discriminators, DR and DA, are used to enhance them, respectively (see Figure 3). During the training process, an image from the realistic domain is randomly selected as the content image I_c and another image from the artistic domain as the style image I_s. I_c and I_s are used as the real samples of D_R and D_A, respectively. The generated image $I_{cs} = G(I_c, I_s)$ is used as the fake sample of D_A. The content and style images are then exchanged to generate an image $I_{sc} = G(I_s, I_c)$ as the fake sample of D_R. The adversarial loss is determined as follows:

$$\begin{aligned} \mathcal{L}_{adv} = \;& \mathbb{E}[\log D_R(I_c)] + \mathbb{E}[\log(1 - D_R(I_{cs}))] \\ & + \mathbb{E}[\log D_A(I_s)] + \mathbb{E}[\log(1 - D_A(I_{sc}))]. \end{aligned} \tag{7}$$

To maintain the content information of the content image in the style transfer between the two domains, a cycle consistency loss is added:

$$\mathcal{L}_{cyc} = \mathbb{E}[\|I_c - G(I_{cs}, I_c)\|_1] + \mathbb{E}[\|I_s - G(I_{sc}, I_s)\|_1]. \tag{8}$$

3.4 Video Style Transfer

To apply our method for video style transfer, the patch-wise contrastive content loss in [Park et al. 2020a] is adopted to keep the content consistency. The feature maps of the content image and the stylized result are cut into feature patches. The patches at the same specific location of the content image and the stylized

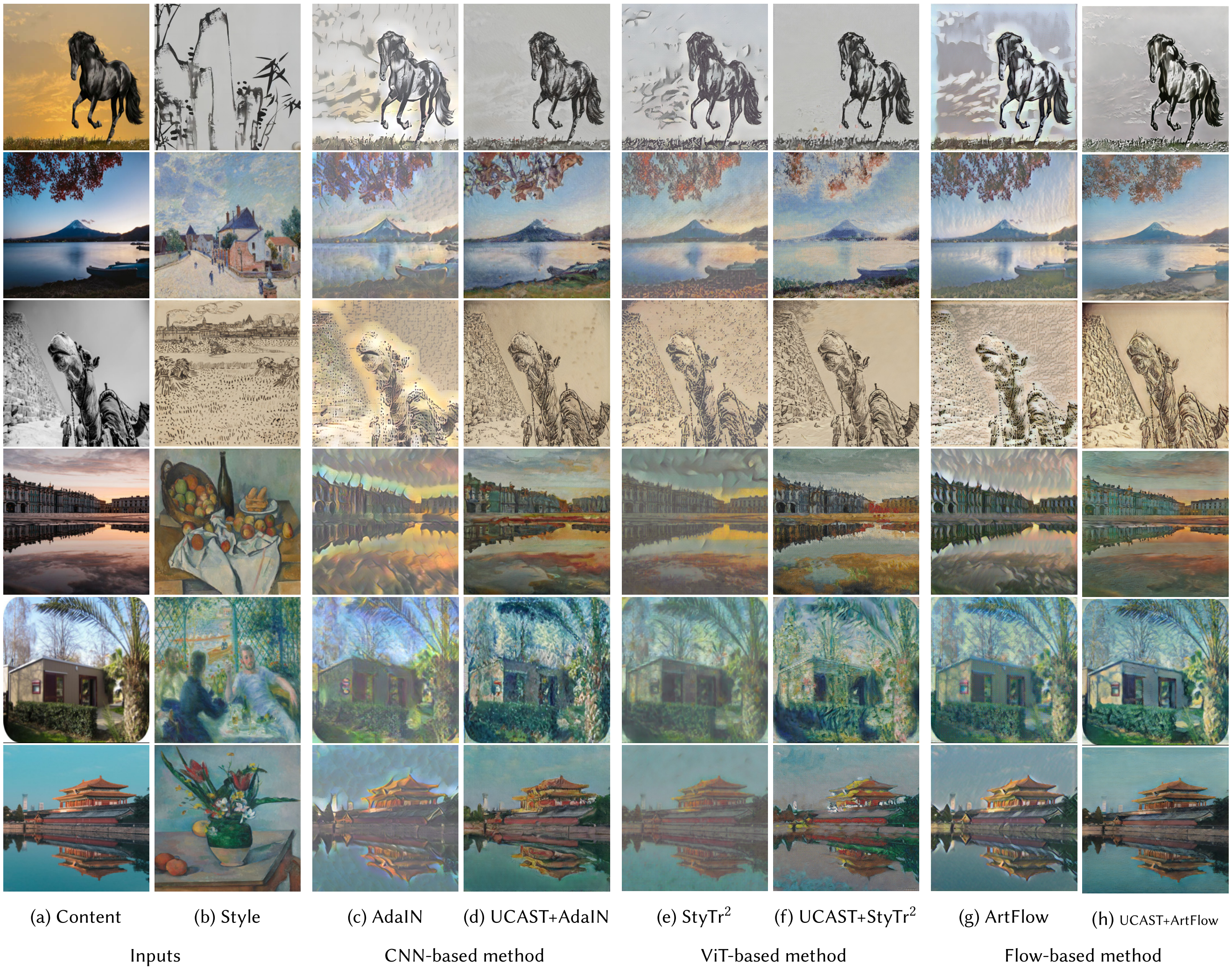


Fig. 6. Qualitative comparisons on different backbones trained under our UCAST framework. Style image (the 2nd row-6th rows) credit: Alfred Sisley/AIC (CC0)[Art Institute of Chicago 2023], Vincent van Gogh/NGA (CC0) [National Gallery of Art 2023], {Paul Cezanne, Pierre-Auguste Renoir, Paul Cezanne}/AIC (CC0)[Art Institute of Chicago 2023].

result are leveraged as positive samples while the other patches within the input are leveraged as negatives:

$$\mathcal{L}^{c}_{contra} = -\log \frac{\exp(v \cdot v^{+}/\tau)}{\exp(v \cdot v^{+}/\tau) + \sum_{n=1}^{W} \exp(v \cdot v_n^{-}/\tau)}, \tag{9}$$

where $v, v^{+} \in \mathbb{R}^{K}$, $v_n^{-} \in \mathbb{R}^{K \times W}$ denote the content feature of the generated image patch, content image patch, and negative image patches, respectively.

3.5 Network Training

Our full objective function for training of the generator G and discriminators D_R and D_A is formulated as follows:

$$\mathcal{L}(G, D_R, D_A) = \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{cyc} + \lambda_3 \mathcal{L}^{G}_{contra} + \lambda_4 \mathcal{L}^{c}_{contra}, \tag{10}$$

where λ_1, λ_2, λ_3, and λ_4 are weights to balance different loss terms. We set $\lambda_1 = 1.0$, $\lambda_2 = 2.0$, $\lambda_3 = 0.2$, and $\lambda_4 = 1.0$ are set in our experiments.

4 EXPERIMENTS

We compare UCAST with several state-of-the-art style transfer methods, including AdaIN [Huang and Belongie 2017], ArtFlow [An et al. 2021], MCCNet [Deng et al. 2021], AdaAttN [Liu et al. 2021b], IEST [Chen et al. 2021a], StyleFormer [Wu et al. 2021a], and StyTr^2 [Deng et al. 2022]. All the baselines are trained using publicly available implementations with default configurations. The comparison of inference speed is shown in Table 1. In all our experiments, our results are generated by using AdaIN as backbone, if no specific annotation is given.

Implementation details. A total of 100,000 artistic images in different styles are collected from WikiArt [Phillips and Mackintosh

Table 1. Statistics of Inference Speed and Quantitative Comparison with State-of-the-art Methods

Method	Inference time	Content loss↓	LPIPS↓	Deception Rate↑	User Study I	User StudyII	
	(ms/image)					Precision↓	Recall↓
StyleTr^2	87	0.123	0.311	54.7%	38.3%	59.0%	56.7%
StyleFormer	**8**	0.176	0.329	53.2%	39.6%	67.2%	63.4%
IEST	184	0.134	0.305	58.7%	41.3%	65.6%	58.6%
AdaAttN	130	0.125	0.304	50.8%	38.9%	63.0%	58.3%
MCCNet	29	0.137	0.308	45.3%	36.2%	73.6%	70.8%
ArtFlow	168	0.121	0.314	44.2%	39.4%	58.8%	55.5%
AdaIN	11	0.160	0.336	51.0%	27.8%	72.4%	64.6%
UCAST+AdaIN	11	**0.117**	0.302	64.2%	-	**39.2**%	**36.3**%
UCAST+StyleTr^2	87	0.122	0.311	**68.2**%	-	-	-
UCAST+ArtFlow	168	0.121	**0.251**	62.0%	-	-	-

The results of user study I represent the average percentage of cases in which the result of the corresponding method is preferred over ours. The results of user study II show the accuracy and recall of being selected as fake paintings by the participants. The best results are in **bold** and the second-best results are underlined.

2011], and 20,000 images are randomly sampled as our artistic dataset. A total of 20,000 images from Places365 [Zhou et al. 2018] are randomly sampled as realistic image dataset. Our framework is trained and evaluated on those artistic and realistic images. In the training phase, all images are loaded with 256×256 resolution. The number of feature map layers M is set to be 4. The dimension K of style latent code is set to 512, 512, 512, and 512 for the four different layers, respectively. Adam [Kingma and Ba 2015] is used as optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a batch size of 4. The initial learning rate is set to 1×10^{-4} and linear decayed linear for total 8×10^5 iterations. The training takes about 18 hours on an NVIDIA GeForce RTX3090.

4.1 Effectiveness on Various Backbones

Our UCAST, as a separate network structure, can be plug-and-play for most arbitrary image style transfer models. In our experiments, UCAST is adapted to AdaIN [Huang and Belongie 2017], ArtFlow [An et al. 2021], and StyTr^2 [Deng et al. 2022]. AdaIN [Huang and Belongie 2017] is a CNN-based style transfer model that includes a fixed VGG network to encode the content and style images, an adaptive instance normalization layer to align the channel-wise mean and variance of content features to match those of style features, and a CNN decoder to invert the AdaIN output to the image spaces. ArtFlow [An et al. 2021] is a neural flow-based model that consists of reversible neural flows and an unbiased feature transfer module. Neural flows are a type of deep generative model that learns the precise likelihood of high-dimensional observations via a series of invertible transformations. StyTr^2 [Deng et al. 2022] is a ViT-based model that contains two transformer encoders for the content image and the style reference, respectively, a multilayer transformer decoder for content sequence stylization, and a CNN decoder.

The comparison results are shown in Figure 6. When transferring style images of ink and wash, as shown in the 1st row, the three backbone methods cannot faithfully generate the brush strokes and the empty background. By training under the UCAST framework, all the enhanced methods can generate high-quality ink and wash images with smooth empty backgrounds and vivid strokes. When dealing with watercolor image, as shown in the 2nd row, the backbones cannot capture the feeling of color blooming. Given that the sky in the content image is a large empty area which the style image does not have, the three backbones tend to generate evident artifacts. Being trained under UCAST can reduce the artifacts and transfer the unique strokes of watercolor. In the 3rd and 4th rows, the backbones fail to transfer the sharp lines in the style reference, whereas UCAST improves the details of the generated images significantly. UCAST can also help all the backbones generate vivid brush strokes of oil paintings, as shown in the 5th row.

4.2 Qualitative Evaluation

4.2.1 Image Style Transfer. First, the qualitative results of our method against the selected state-of-the-art methods are presented in Figure 7. The comparison shows the superiority of UCAST in terms of visual quality. AdaIN often fails to generate sharp details and introduces undesired patterns that do not exist in style images (e.g., the 4th, 6th, 9th, and 11th rows). ArtFlow sometimes generates unexpected colors or patterns in relatively smooth regions in some cases (e.g., the 2nd, 3rd, and 8th rows). MCCNet can effectively preserve the input content but may fail to capture the stroke details and often generates haloing artifacts around object contours (e.g., the 2nd, 5th, 9th rows). AdaAttN cannot well capture some stroke patterns and fails to transfer important colors of the style references to the results (e.g., the 1st, 5th, and 6th rows). Although the generated visual effects of IEST are of high quality, the usage of second-order statistics as style representation causes color distortion (e.g., the 1st and the 4th row) and cannot capture the detailed stylized patterns (e.g., the 5th and 7th rows). StyleFormer cannot well capture some stroke patterns and tends to generate artifacts in the results (e.g., the 1st, 6th, and 8th rows). StyTr^2 cannot well transfer the unique style of the reference images and also tends to generate artifacts(e.g., the 1st, 3rd, and 4th rows). In particular, these state-of-the-art methods cannot capture the *leaving blank* characteristic of the Chinese painting style in the 1st row of Figure 7 and fail to generate results with a clean background.

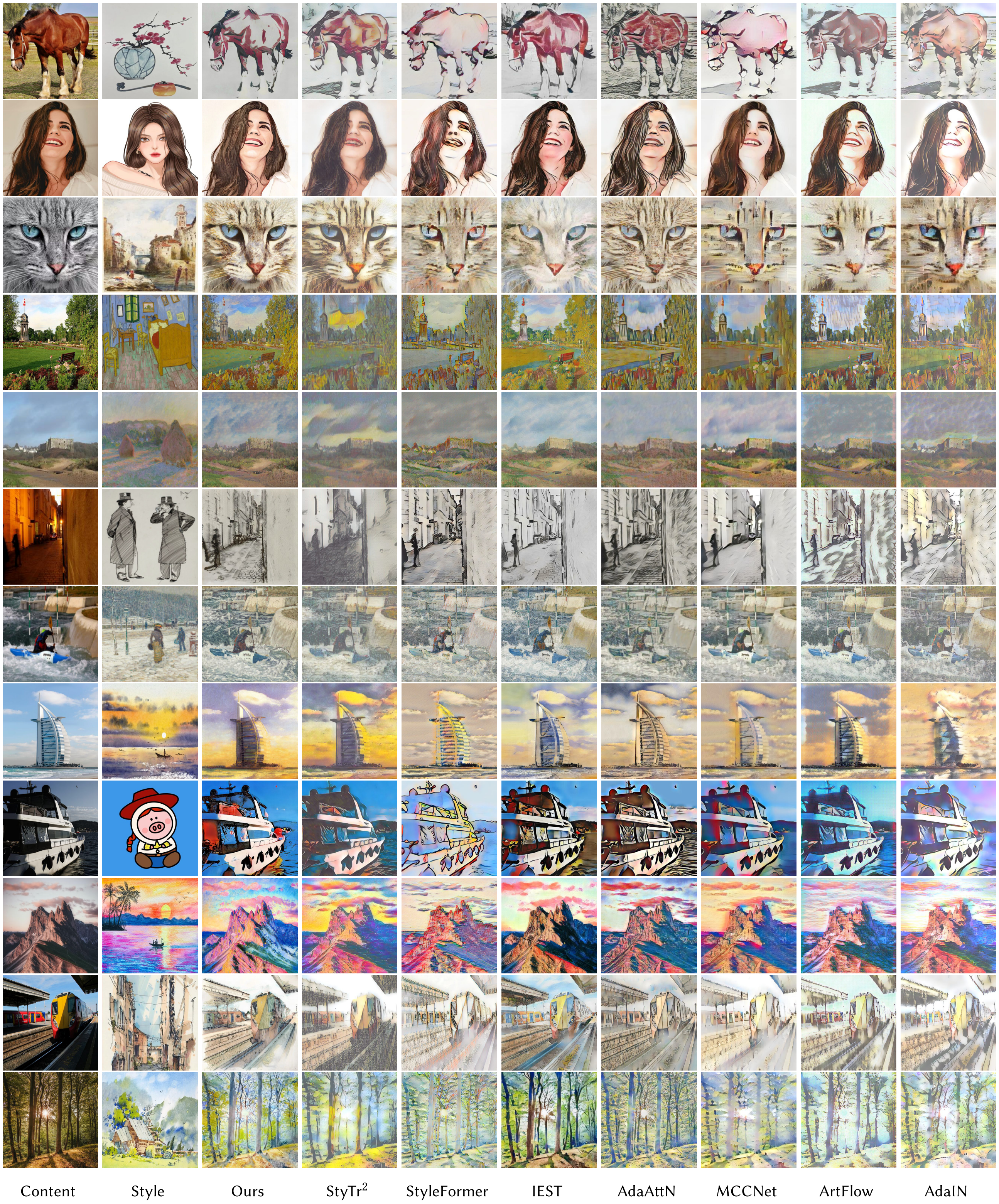


Fig. 7. Qualitative comparisons with several state-of-the-art style transfer methods, including StyTr^2 [Deng et al. 2022], StyleFormer [Wu et al. 2021a], IEST [Chen et al. 2021a], AdaAttN [Liu et al. 2021b], MCCNet [Deng et al. 2021], ArtFlow [An et al. 2021], AdaIN [Huang and Belongie 2017]. Content image credits (the 1st–3rd rows): {Pixabay, Thaís Sarmento, Pixabay}/Pexels (Free to use) [Pexels 2023]. Style image credits (the 4th–7th rows): {Vincent van Gogh, Claude Monet, Philip William May, Childe Hassam}/AIC (CC0) [Art Institute of Chicago 2023].

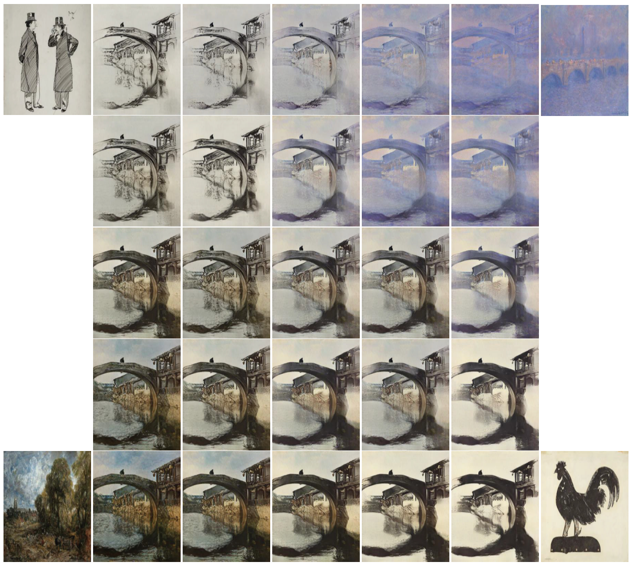


Fig. 8. Linear interpolation results of multiple styles. The input style images are shown in the four corners. Image credits (from left to right, from top to bottom): {Philip William May, Claude Monet, John Constable, Childe Hassam}/AIC (CC0) [Art Institute of Chicago 2023], Nicholas Acampora/NGA (CC0) [National Gallery of Art 2023].

In comparison, UCAST achieves the best stylization performance that balances the characteristics of style patterns and content structures. Instead of using second-order statistics as a global style descriptor, an MSP module is used for style encoding with the help of a DE module for effective learning of style distribution. Thus, UCAST can flexibly represent vivid local stroke characteristics and the overall appearance while preserving the content structure. For instance, as shown in Figures 1, 2(c) (the 1st row), and 7 (the 1st row), UCAST successfully captures the large portion of empty regions in the style images, and it generates a stylization results that have salient objects in the center and blank space around. As shown in Figure 7, besides commonly used oil paintings (the 2nd, 3rd, and 5th rows), UCAST can also generate high-quality results of line drawing (the 2nd row), cartoon (the 7th–9th rows), aquarelle (the 6th and 11th rows), crayon drawing (the 10th row) and color pencil drawing (the 12th row).

4.2.2 Style Interpolation. The feature maps among four style images with equivalent weights are interpolated. Figure 8 shows interpolation can be done among arbitrary styles by providing the decoder with a convex mixture of feature maps converted to various styles. Smooth intra-domain (vertically) and inter-domain (horizontally) interpolation results are obtained.

4.3 Quantitative Evaluation

The content loss [Li et al. 2017], LPIPS [Chen et al. 2021a], and deception rate [Sanakoyeu et al. 2018b] are used, and two user studies are conducted to evaluate our method quantitatively. The two user studies are online surveys that cover art/computer science students/professors and civil servants.

For content loss and LPIPS, a pre-trained VGG-19 is used, and the average perceptual distances between the content image and the stylized image are computed. The statistics are shown in Table 1. For deception rate, a VGG-19 network is trained to classify ten styles on WikiArt. Then, the deception rate is calculated as the percentage of stylized images predicted by the pre-trained network as the correct target styles. The deception rate for the proposed UCAST and the baseline models are reported in the 2nd column of Table 1. As observed, UCAST achieves the highest accuracy and surpasses other methods by a large margin. As a reference, the mean accuracy of the network on real images of the artists from WikiArt is 78%.

User Study I. We compare UCAST with seven state-of-the-art style transfer methods to evaluate which method generates results that are most favored by humans. For each participant, 50 content-style pairs are randomly selected, and in each question, the stylized result of UCAST and one of the other methods are displayed in random order. Firstly, the purpose of the style transfer task is introduced to the participants, i.e., transferring the style of a painting image to a photo to generate a picture with corresponding content and style. For each question, the participant is asked to choose the better image that learns the most characteristics from the style image and maintains the semantic information of the content image. There is neither a training period nor specific guidelines (e.g., the definition of the “characteristics”) given that most of the participants are familiar with image synthesis or art analysis. In this manner, the faithful preference results of professionals can be obtained. Finally, 3,800 votes are collected from 76 participants (52 computer graphics or computer vision researchers, 12 artists, and other 12 people with different backgrounds). The percentage of votes for each method is reported in the 6th column of Table 1. These results demonstrate that UCAST achieves better style transfer results. Moreover, according to the statistics, UCAST obtains significantly higher preferences in categories of sketch, Chinese painting, and impressionism.

User Study II. A novel user study is designed to evaluate the stylized images quantitatively, which is called the Stylized Authenticity Detection. For each question, participants are shown ten artworks of similar styles, including two to four stylized fake paintings, and asked to select the synthetic ones. Within each single question, the stylized paintings are generated by the same method. Each participant finishes 25 questions. Finally, we collect 2,000 groups of results collected from 80 participants (55 computer graphics or computer vision researchers, 12 artists, and other 13 people with different backgrounds), and the average precision and recall are used as the measurement for how likely the results are recognized as synthetics. The percentage of votes for each method is reported in the 7th column of Table 1. The paintings generated by UCAST have the lowest chance to be decided by people as fake paintings. Moreover, the precision and recall of UCAST are less than 50%, which means users could not distinguish the real ones from the fakes, and they select more real paintings as synthetics during the testing.

4.4 Video Style Transfer

We compare our method with seven baselines on video style transfer and show the stylization results in Figure 9. The heat maps of differences between different frames are visualized to

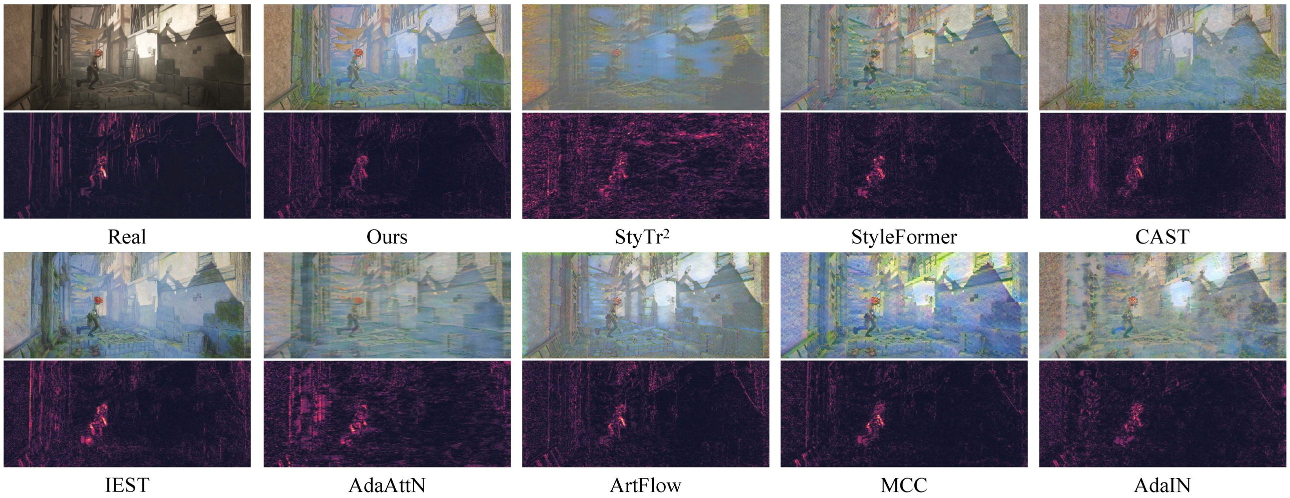


Fig. 9. Qualitative comparison on video style transfer. The first column shows the input video frame and the rest of the columns show the stylization results generated by different style transfer methods. The heat map of differences between the current frame and the previous adjacent frame are shown beneath each frame.

Table 2. Quantitative Evaluation of Temporal Consistency on 50 Rendered Clips

	Ours	StyTr^2	StyleFormer	CAST	IEST	AdaAttN	ArtFlow	MCCNet	AdaIN
Temporal Loss↓	**0.0322**	0.0350	0.0352	0.0439	0.0460	0.0367	0.0329	0.0441	0.0489

The best results are in **bold**.

assess the stability and consistency of synthesized video clips. Our approach outperforms existing style transfer methods in terms of stability and consistency by a significant margin. This result can be attributed to three points: (1) Our style representation and domain distribution learning offer proper guidance to prevent the model from distorted texture patterns. (2) The cycle consistency loss enhances the consistency of the synthesized video clip. (3) The added patch-wise contrastive losses offer a strong content consistency constraint which motivates the same object in a different frame to have the same stylization results.

Video consistency. The widely used temporal loss [Wang et al. 2020] is employed to quantitatively analyze the temporal consistency of stylized videos. Given two adjacent frames I_c^t and I_c^{t-1} in a T-frame input clip and I_{cs}^t and I_{cs}^{t-1} in a T-frame rendered clip, the temporal loss is defined as follows:

$$L_{temporal} = average(||O \circ (W_{I_c^{t-1} \to I_c^t}(I_{cs}^{t-1}) - I_{cs}^t)||), \tag{11}$$

where O is an occlusion mask:

$$O = |W_{I_c^{t-1} \to I_c^t}(I_{cs}^{t-1}) - I_{cs}^t| > 10. \tag{12}$$

Table 2 shows our method achieves the best temporal consistency.

4.5 Ablation Study

Contrastive style loss. We remove the contrastive style loss from Equation (10) to train the model. As shown in Figure 10(b), the model without our contrastive style loss cannot capture the color and the stroke characteristics of the style image compared with the full model. The brushstrokes of watercolor in the style image almost disappear in the 1st row. The sharp lines and edges in the 2nd row become smooth and murky. The brown color of the whole image generated in the 3rd row does not appear in the style image.

We replace the adaptive temperature from Equation (3) with constant temperature to train the model. Figure 10(c) shows that when dealing with difficult content-style pairs, the model without our adaptive temperature tends to generate artifacts. For instance, the black artifact appears in the sky of the 1st row and 3rd row. By introducing input-dependent temperature, the full UCAST can capture and transfer the unique style of cartoon. In the 2nd row, the sharp lines and flat color fillings in the style image are faithfully transferred to the results while the simplified model generates result with mixing style. The content details of the women's face are well preserved by the full model. With the contrastive style loss and adaptive temperature, our full model can faithfully transfer the brushstrokes, textures, and colors from the input style image.

Domain enhancement. Our full UCAST uses DE for realistic and artistic images separately. We train a simplified UCAST model without DE module. Figure 11(d) and (h) shows the color of the style images are faithfully transferred, but the generated images do not appear like real paintings. A simplified UCAST model is trained using one discriminator that mixes realistic and artistic images together (mix-DE). Figure 11(e) shows the results generated by the mix-DE model are acceptable, but the stroke details in the generated images are weaker than those ones by the full UCAST model. This fact is due to the existence of a significant gap between the artistic and realistic image domains. All images

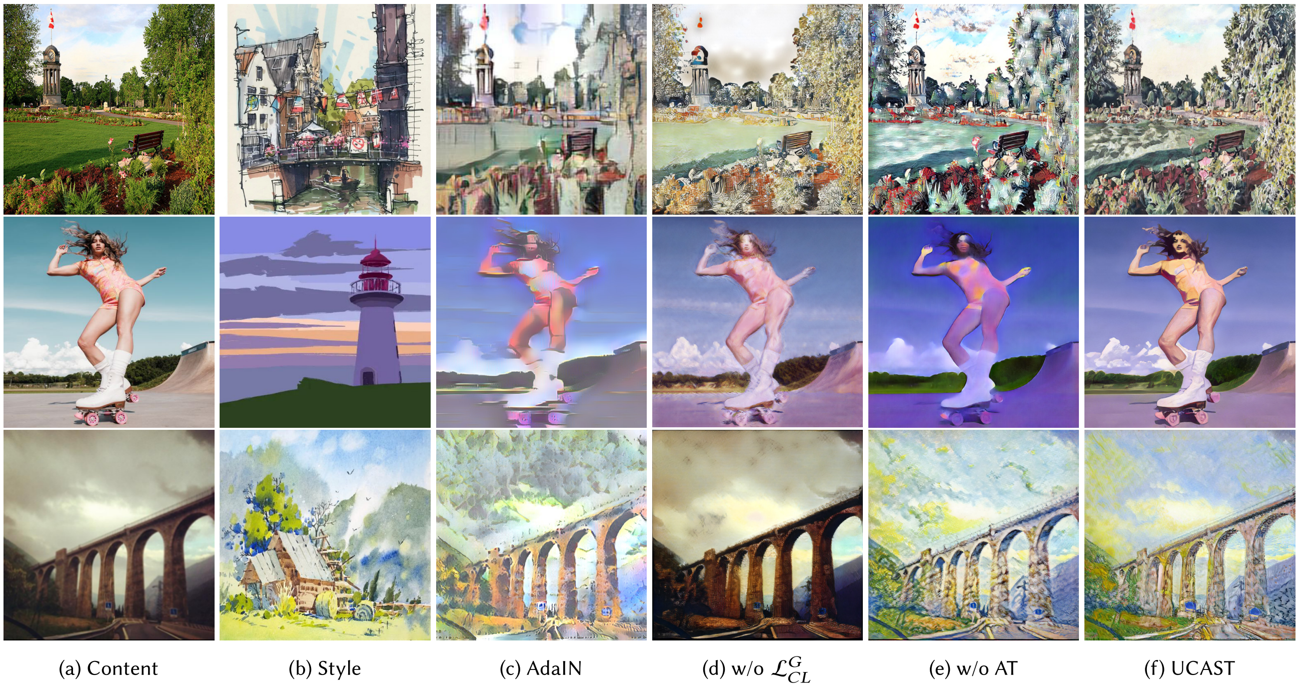


Fig. 10. Ablation study on adaptive contrastive learning. From left to right: (a) content image; (b) style image; (c) AdaIN; (d) UCAST without contrastive loss; (e) UCAST without adaptive temperature; (f) full UCAST. Content image credit (the 2nd row): Airam Dato-on/Pexels (Free to use) [Pexels 2023].

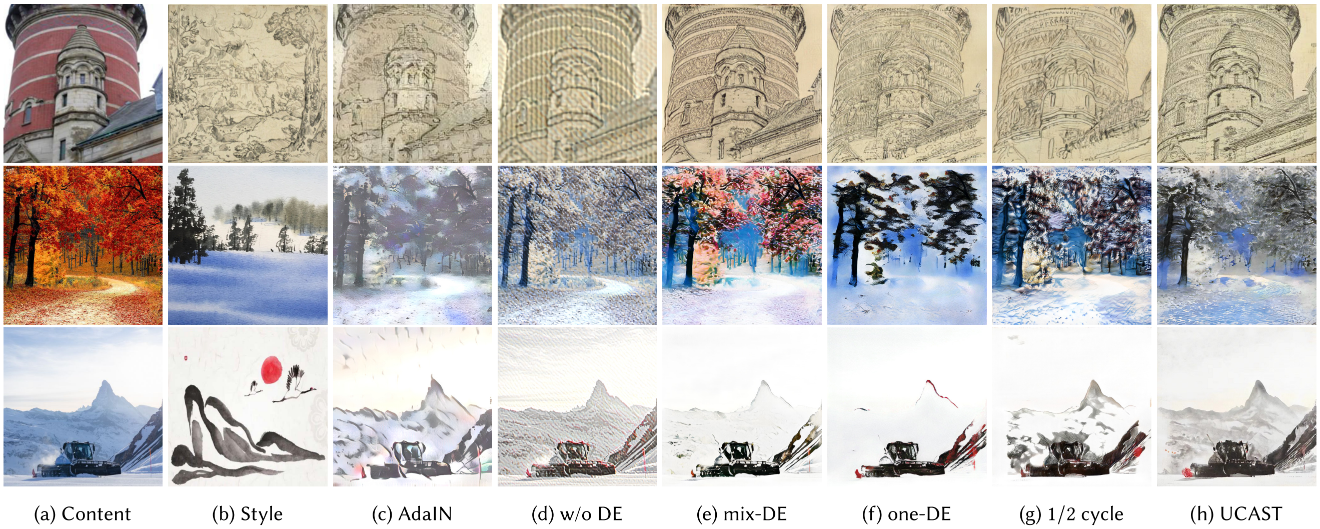


Fig. 11. Ablation study on DE. From left to right: (a) content image; (b) style reference; (c) AdaIN; (d) UCAST without DE; (e) UCAST using mixed DE; (f) UCAST using one DE without the realistic domain; (g) UCAST trained with the asymmetric cycle consistent loss by only reconstructing the realistic images; and (h) the full UCAST model. Style image (1st row) credits: Michel Ange Corneille/AIC(CC0) [Art Institute of Chicago 2023].

from the realistic domain are abandoned for ablation (one-DE). As shown in Figure 11(f), the results generated by the one-DE model lack details.

To better evaluate the improvement of the contrastive style loss on the style transfer task, the latent promotion of cycle consistency loss is excluded from network training because the reconstruction of artistic image may imply style information. UCAST is trained with an asymmetric cycle consistent loss, which only reconstructs the realistic images. The decoder of the style transfer network is unaffected by the reconstruction of the artistic image. Figure 11(g) shows that removing realistic image reconstruction will lead to slightly degraded stylization results.

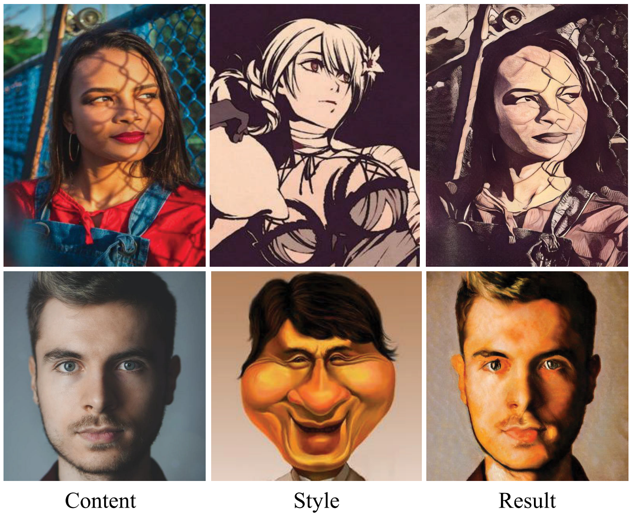


Fig. 12. Typical failure cases of UCAST. Content image credits: {David Gomes, Simon Robben} (Free to use)/Pexels [Pexels 2023].

4.6 Limitations

UCAST has limited capability in the fine-grained controllability of specific objects. If an object in the style image is in a specific color, sometimes it fails to transfer the color in a semantic matching way. For example, in the first row of Figure 12, the red color of the eyes in the style image is not transferred to the eyes in the content image but appears in some regions of the clothes in the generated image. A possible improvement would be to analyze the semantic information represented by different dimensions of the style code to enhance the controllability of the model. UCAST also has difficulty to producing large geometric change, like the example shown in the second row of Figure 12, where UCAST fails to transfer the special face shape in the style image to the content image.

5 CONCLUSION AND FUTURE WORK

In this work, a novel unified framework, namely, UCAST, is presented for the task of arbitrary image style transfer. Instead of relying on second-order metrics such as Gram matrix or mean/variance of deep features, image features are used directly by introducing an MSP module for style encoding. A parallel contrastive learning scheme is developed to leverage the available multistyle information in the existing collection of artwork and help train the MSP module and the generative style transfer network. An adaptive contrastive learning is proposed for style transfer implemented by a dual input-dependent temperature. A DE scheme is further suggested to effectively model the distribution of realistic and artistic image domains. The extensive experimental results demonstrate our proposed UCAST method is effective for various generative backbones and achieves superior arbitrary style transfer results compared with state-of-the-art approaches. In the future, the contrastive style learning will be improved by considering artist and category information.

REFERENCES

Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. 2021. ArtFlow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conferences on Computer Vision and Pattern Recognition (CVPR).* 862–871.

Art Institute of Chicago. 2023. (2023). Retrieved June 03, 2023 from https://www.artic.edu/.

Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. 2021. Rethinking the truly unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* 14154–14163.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* 9650–9660.

Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. 2017. StyleBank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 1897–1906.

Haibo Chen, Lei Zhao, Zhizhong Wang, Zhang Hui Ming, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2021a. Artistic style transfer with internal-external learning and contrastive learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS).*

Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2021b. DualAST: Dual style-learning networks for artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 872–881.

Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. 2021. Arbitrary video style transfer via multi-channel correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI).* 1210–1217.

Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. 2022. StyTr^2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 11326–11336.

Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. 2020. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the ACM International Conference on Multimedia.* 2719–2727.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2017. A learned representation for artistic style. In *Proceedings of the International Conference on Learning Representations.*

Jakub Fišer, Ondřej Jamriška, Michal Lukáč, Eli Shechtman, Paul Asente, Jingwan Lu, and Daniel Sýkora. 2016. StyLit: Illumination-guided example-based stylization of 3D renderings. *ACM Transactions on Graphics* 35, 4 (2016), 11 pages.

Wei Gao, Yijun Li, Yihang Yin, and Ming-Hsuan Yang. 2020. Fast video multi-style transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).* 3222–3230.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2414–2423.

Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. 2017. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 3730–3738.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS).*

Junlin Han, Mehrdad Shoeiby, Lars Petersson, and Mohammad Ali Armin. 2021. Dual contrastive learning for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.* 746–755.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 9729–9738.

Nisha Huang, Fan Tang, Weiming Dong, and Changsheng Xu. 2022a. Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion. In *Proceedings of the 30th ACM International Conference on Multimedia.* 1085–1094.

Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Yong Zhang, Weiming Dong, and Changsheng Xu. 2022b. DiffStyler: Controllable dual diffusion for text-driven image stylization. arXiv:2211.10682. Retrieved from https://arxiv.org/abs/2211.10682.

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 1501–1510.

Jongheon Jeong and Jinwoo Shin. 2021. Training GANs with stronger augmentations via contrastive discriminator. In *Proceedings of the International Conference on Learning Representations.*

Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. 2020a. Dynamic instance normalization for arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 4369–4376.

Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. 2020b. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics* 26, 11 (2020), 3365–3385.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV).* Springer, 694–711.

Minguk Kang and Jaesik Park. 2020. ContraGAN: Contrastive learning for conditional image generation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS).*

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR).*

Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. 2019. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 10043–10052.

Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. 2019a. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* 4422–4431.

Dmytro Kotovenko, Artsiom Sanakoyeu, Pingchuan Ma, Sabine Lang, and Bjorn Ommer. 2019b. A content transformation block for image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 10032–10041.

Gihyun Kwon and Jong Chul Ye. 2022. CLIPstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 18062–18071.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV).* 35–51.

Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. 2019. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 3804–3812.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. In *Proceedings of the Advances Neural Information Processing Systems (NeurIPS).* 386–396.

Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. 2017. Visual attribute transfer through deep image analogy. *ACM Transactions on Graphics* 36, 4 (2017), 15 pages.

Minxuan Lin, Fan Tang, Weiming Dong, Xiao Li, Changsheng Xu, and Chongyang Ma. 2021. Distribution aligned multimodal and multi-domain image stylization. *ACM Transactions on Multimedia Computing, Communications, and Applications* 17, 3 (2021), 17 pages.

Rui Liu, Yixiao Ge, Ching Lam Choi, Xiaogang Wang, and Hongsheng Li. 2021a. DivCo: Diverse conditional image synthesis via contrastive generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 16372–16381.

Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. 2021b. AdaAttN: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* 6649–6658.

Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. 2019. Exploiting unlabeled data in CNNs by self-supervised learning to rank. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 8 (2019), 1862–1878.

Thaneeya McArdle. 2022. Explore art styles. (2022). Retrieved from https://www.art-is-fun.com/art-styles. Accessed 3 October 2022.

National Gallery of Art. 2023. (2023). Retrieved June 03, 2023 from https://www.nga.gov/.

Dae Young Park and Kwang Hee Lee. 2019. Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 5880–5888.

Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. 2020a. Contrastive learning for unpaired image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV).* Springer, 319–345.

Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. 2020. Swapping autoencoder for deep image manipulation. In *Proceedings of the Advances Neural Information Processing Systems (NeurIPS).* 7198–7211.

Pexels. 2023. (2023). Retrieved June 03, 2023 from https://www.pexels.com.

Fred Phillips and Brandy Mackintosh. 2011. Wiki Art Gallery, Inc.: A case for critical thinking. *Issues in Accounting Education* 26, 3 (2011), 593–608.

Gilles Puy and Patrick Pérez. 2019. A flexible convolutional solver for fast style transfers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 8963–8972.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 10684–10695.

Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Bjorn Ommer. 2018a. A style-aware content loss for real-time HD style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV).* 698–714.

Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Björn Ommer. 2018b. A style-aware content loss for real-time HD style transfer. In *Proceedings of the European Conference Computer Vision (ECCV).* Springer International Publishing, Cham, 715–731.

Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, and Stephen Gould. 2019. Visual permutation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 12 (2019), 3100–3114.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR).*

Jan Svoboda, Asha Anoosheh, Christian Osendorfer, and Jonathan Masci. 2020. Two-stage peer-regularized feature recombination for arbitrary image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 13816–13825.

Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. 2016. Texture networks: Feed-forward synthesis of textures and stylized images. In *Proceedings of the International Conference on Machine Learning (ICML).* 1349–1357.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2019).

Bin Wang, Wenping Wang, Huaiping Yang, and Jiaguang Sun. 2004. Efficient example-based painting and synthesis of 2D directional texture. *IEEE Transactions on Visualization and Computer Graphics* 10, 3 (2004), 266–277.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2495–2504.

Qian Wang, Cai Guo, Hong-Ning Dai, and Ping Li. 2023. Stroke-GAN painter: Learning to paint artworks using stroke-style generative adversarial networks. *Computational Visual Media* (2023). https://link.springer.com/article/10.1007/s41095-022-0287-3.

Wenjing Wang, Jizheng Xu, Li Zhang, Yue Wang, and Jiaying Liu. 2020. Consistent video style transfer via compound regularization. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI).* AAAI, 12233–12240.

Zhizhong Wang, Zhanjie Zhang, Lei Zhao, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2022. AesUST: Towards aesthetic-enhanced universal style transfer. In *Proceedings of the 30th ACM International Conference on Multimedia.* 1095–1106.

Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. 2021a. StyleFormer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* 14598–14607.

Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. 2021b. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 10546–10555.

Zijie Wu, Zhen Zhu, Junping Du, and Xiang Bai. 2022. CCPL: Contrastive coherence preserving loss for versatile style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV).* Springer, 189–206.

Wenju Xu, Chengjiang Long, Ruisheng Wang, and Guanghui Wang. 2021. DRB-GAN: A dynamic resblock generative adversarial network for artistic style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* 6383–6392.

Chaoning Zhang, Kang Zhang, Trung X. Pham, Axi Niu, Zhinan Qiao, Chang D. Yoo, and In So Kweon. 2022a. Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 14441–14450.

Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Trung X. Pham, Chang D. Yoo, and In So Kweon. 2022b. How does SimSiam avoid collapse without negative samples? A unified understanding with self-supervised contrastive learning. In *Proceedings of the International Conference on Learning Representations (ICLR).*

Hang Zhang and Kristin Dana. 2018. Multi-style generative network for real-time transfer. In *Proceedings of the European Conference on Computer Vision Workshops.* 349–365.

Oliver Zhang, Mike Wu, Jasmine Bayrooti, and Noah Goodman. 2021. Temperature as uncertainty in contrastive learning. In *Proceedings of the NeurIPS Self-Supervised Learning—Theory and Practice Workshop.*

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023a. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 10146–10156.

Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. 2022a. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8035–8045.

Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. 2022b. Domain enhanced arbitrary image style transfer via contrastive learning. In *Proceedings of the 2022 Conference on ACM SIGGRAPH* . 8 pages.

Yuxin Zhang, Fan Tang, Weiming Dong, Thi-Ngoc-Hanh Le, Changsheng Xu, and Tong-Yee Lee. 2023bb. Portrait map art generation by asymmetric image-to-image translation. *Leonardo* 56, 1 (2023), 28–36.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 6 (2018), 1452–1464.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2223–2232.

Received 19 October 2022; revised 7 May 2023; accepted 24 May 2023