MicroAST: Towards Super-fast Ultra-Resolution Arbitrary Style Transfer

Zhizhong Wang, Lei Zhao*, **Zhiwen Zuo, Ailin Li,**
Haibo Chen, Wei Xing*, **Dongming Lu**
College of Computer Science and Technology, Zhejiang University
{endywon, cszhl, zzwcs, liailin, cshbchen, wxing, ldm}@zju.edu.cn

Abstract

Arbitrary style transfer (AST) transfers arbitrary artistic styles onto content images. Despite the recent rapid progress, existing AST methods are either incapable or too slow to run at ultra-resolutions (e.g., 4K) with limited resources, which heavily hinders their further applications. In this paper, we tackle this dilemma by learning a straightforward and lightweight model, dubbed MicroAST. The key insight is to completely abandon the use of cumbersome pre-trained Deep Convolutional Neural Networks (e.g., VGG) at inference. Instead, we design two micro encoders (content and style encoders) and one micro decoder for style transfer. The content encoder aims at extracting the main structure of the content image. The style encoder, coupled with a modulator, encodes the style image into learnable dual-modulation signals that modulate both intermediate features and convolutional filters of the decoder, thus injecting more sophisticated and flexible style signals to guide the stylizations. In addition, to boost the ability of the style encoder to extract more distinct and representative style signals, we also introduce a new style signal contrastive loss in our model. Compared to the state of the art, our MicroAST not only produces visually superior results but also is 5-73 times smaller and 6-18 times faster, for the first time enabling super-fast (about 0.5 seconds) AST at 4K ultra-resolutions.

1 Introduction

Style transfer has recently attracted ever-growing interest in both academia and industry since the seminal work of (Gatys, Ecker, and Bethge 2016). A central problem in this domain is the task of transferring the artistic style of an arbitrary image onto a content target, which is called *arbitrary style transfer (AST)* (Huang and Belongie 2017; Li et al. 2017). By leveraging the remarkable representative power of pre-trained Deep Convolutional Neural Networks (DCNNs) (*e.g.*, VGG-19 (Simonyan and Zisserman 2014)), existing AST algorithms consistently achieve both stunning stylizations and generalization ability on arbitrary images. However, the large pre-trained DCNNs incur a high computational cost, which impedes the current AST methods to process ultra-high resolution (*e.g.*, "4K" or 4096×2160 pixels) images with limited resources. It heavily restricts their further applications in practical scenes, such as large posters, ultra high-definition (UHD) photographs, and UHD videos.

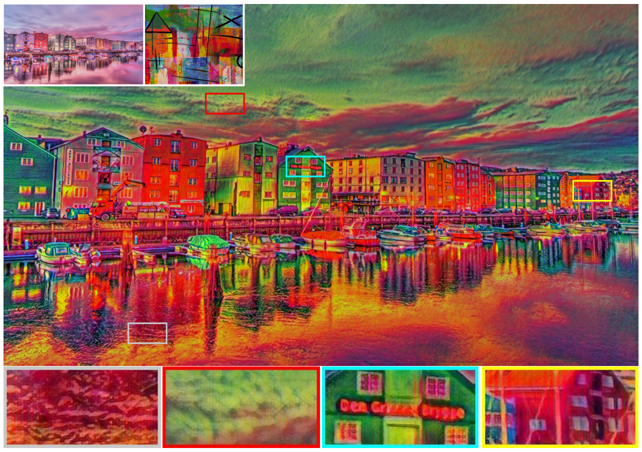


Figure 1: A 4K (4096×2160 pixels) ultra-resolution stylized result, rendered in about 0.5 seconds by our proposed MicroAST on a single NVIDIA RTX 2080 (8GB) GPU. On the upper left are the content and style images. Four close-ups (256×128 pixels) are shown under the stylized image.

Valuable efforts have been devoted to solving this dilemma. One practice is to compress the large pre-trained DCNN models without losing much performance. (Wang et al. 2020a) used collaborative distillation to reduce the convolutional filters of VGG-19, successfully rendering ultra-resolution images on a single 12GB GPU. While the memory consumption is significantly reduced, the pruned models are often not fast enough to run at ultra-resolutions. Another solution is to stylize the images in a patch-wise manner (Chen et al. 2022). This method, though achieving unconstrained resolution style transfer, still suffers from the efficiency problem. Similar to our method, (Shen, Yan, and Zeng 2018) and (Jing et al. 2020) likewise designed lightweight networks for style transfer. However, since their style features are still extracted from VGG, they are inherently difficult to process ultra-resolution images. Therefore, despite the recent progress, existing AST methods are still incapable or too slow to run at ultra-resolutions.

Facing the challenges above, in this paper, we propose a

*Corresponding authors.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

straightforward and lightweight model for *super-fast* ultra-resolution arbitrary style transfer. The key insight is that we completely abandon the use of cumbersome pre-trained DCNNs (*e.g.*, VGG) *at inference*, whether for content extraction (Huang and Belongie 2017; Li et al. 2017), or style extraction (Shen, Yan, and Zeng 2018; Jing et al. 2020). Our model, dubbed *MicroAST*, uses two micro encoders and one micro decoder for style transfer. The micro encoders consist of a content encoder and a style encoder. The content encoder aims at extracting the main structure of the content image. The style encoder, coupled with a modulator, encodes the style image into learnable *dual-modulation* signals that modulate both intermediate features and convolutional filters of the decoder. This novel dual-modulation strategy injects more sophisticated and flexible style signals to guide the stylizations, thus helping our model fully capture the global attributes and local brushstrokes of the artistic styles. The decoder generates the final stylized images under these modulations. In addition, to boost the ability of the style encoder to extract more distinct and representative modulation signals for each style, we also introduce a new *style signal contrastive loss* in our model, which further improves the quality. Comprehensive experiments have been conducted to demonstrate the effectiveness of our method. Compared to the state of the art, our MicroAST not only produces visually superior results but also is 5-73 times smaller and 6-18 times faster, for the first time enabling *super-fast* (about 0.5 seconds) AST on 4K ultra-resolution images (see an example in Fig. 1).

In summary, our contributions are threefold:

- We propose a straightforward and lightweight framework called *MicroAST* to achieve *super-fast* ultra-resolution arbitrary style transfer for the first time.
- We introduce a novel *dual-modulation* strategy to inject more sophisticated and flexible style signals to guide the stylizations in our model.
- We also introduce a new *style signal contrastive loss* to boost the ability of our style encoder.

2 Related Work

Neural Style Transfer. The seminal work of (Gatys, Ecker, and Bethge 2016) has opened up the era of Neural Style Transfer (NST) (Jing et al. 2019). In their work, the artistic style of an image is captured by the correlations between features extracted from a pre-trained DCNN. It is amazingly effective and has inspired a lot of successors to improve the performance in many aspects, including efficiency (Johnson, Alahi, and Fei-Fei 2016; Ulyanov et al. 2016), quality (Jing et al. 2018, 2022; Wang et al. 2020c, 2021, 2022a; Lin et al. 2021; Cheng et al. 2021; An et al. 2021; Liu et al. 2021b; Chen et al. 2021b,a; Deng et al. 2020, 2021, 2022; Lu and Wang 2022; Xie et al. 2022), generalization (Huang and Belongie 2017; Li et al. 2017; Sheng et al. 2018; Li et al. 2019; Park and Lee 2019; Lu et al. 2019; Zhang, Zhu, and Zhu 2019; Chiu 2019; Hong et al. 2021; Zhang et al. 2022a), diversity (Wang et al. 2020b, 2022c; Chen et al. 2020a, 2021c), and user control (Champandard 2016; Wang et al. 2022b; Zuo et al. 2022). Despite the monumental progress, existing NST methods all share a fundamental flaw, *i.e.*, they are unable to process ultra-resolution (*e.g.*, 4K) images with limited resources, since they all heavily rely on the large DCNN models (*e.g.*, VGG-19 (Simonyan and Zisserman 2014)) to extract representative features.

Ultra-Resolution Style Transfer. To address the challenges above, (Wang et al. 2020a) used model compression (called collaborative distillation) to reduce the convolutional filters of VGG-19, firstly rendering ultra-resolution images on a 12GB GPU. While the memory consumption is significantly reduced, the pruned models are still not fast enough to run at ultra-resolutions. Besides, a large degree of compression often leads to severe quality degradation.

Another solution is to design a lightweight model directly. (Johnson, Alahi, and Fei-Fei 2016) and (Sanakoyeu et al. 2018) learned small feed-forward networks for a specific style example or category for high-resolution (*e.g.*, 1024×1024 pixels) style transfer. However, they are not generalized to other unseen styles and not capable of running at ultra-resolutions. (Shen, Yan, and Zeng 2018) and (Jing et al. 2020) designed lightweight networks for AST, but they still used pre-trained VGG to extract style features, leading to the expensive cost of extra memory and slow inference speed. Unlike these methods, our approach completely gets rid of the high-cost pre-trained VGG *at inference*, for the first time achieving *super-fast* ultra-resolution style transfer for arbitrary styles with one model only.

Recently, (Chen et al. 2022) provided a possible solution for unconstrained resolution style transfer. They divided input images into small patches and performed patch-wise stylization with a Thumbnail Instance Normalization to ensure the style consistency among different patches. However, this method does not consider the time cost problem and cannot achieve *super-fast* ultra-resolution style transfer.

Contrastive Learning. Contrastive learning has been widely used in self-supervised representation learning for high-level vision tasks (He et al. 2020; Chen et al. 2020b; Tian, Krishnan, and Isola 2020). Recently, in low-level generative tasks, some works investigate the use of contrastive loss for different objectives, such as image-to-image translation (Park et al. 2020), image generation (Kang and Park 2020; Liu et al. 2021a), image dehazing (Wu et al. 2021), and style transfer (Chen et al. 2021a; Zhang et al. 2022b; Wu et al. 2022), *etc*. Unlike these works, we introduce a novel mini-batch *style signal contrastive loss* to help address the problem of ultra-resolution style transfer, which considers relations between multiple style modulation signals in the same training mini-batches. Therefore, it can significantly boost the ability of the micro style encoder to extract more distinct and representative style signals.

3 Proposed Approach

Given *arbitrary* ultra-resolution (*e.g.*, 4K) content image C and style image S, our goal is to produce the corresponding ultra-resolution stylized output O in a *very short* time (e.g., within one second). The challenges mainly lie in three aspects. (i) The method should be capable of processing ultra-resolution images with *limited resources* (e.g., on an 8GB

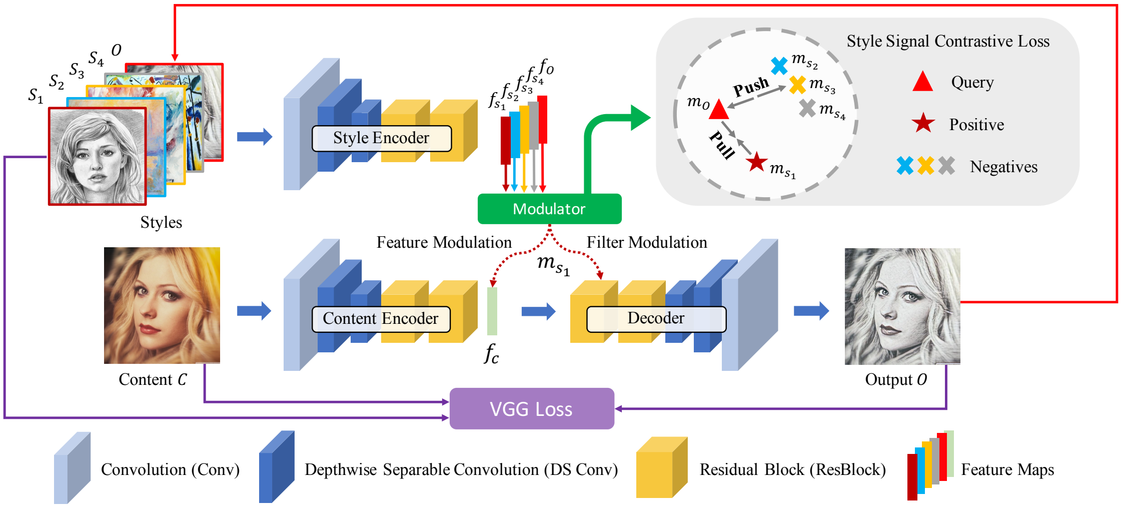


Figure 2: Overview of MicroAST. C denotes the content image. S_i denotes the i^{th} style image in a training mini-batch (we put four style images for illustration). O denotes the stylized output. $f_{\{\dots\}}$ denote the encoded feature maps. $m_{\{\dots\}}$ denote the style modulation signals.

GPU). (ii) The method should generate the ultra-resolution results at a *super-fast* speed. (iii) The method should be able to produce pleasing stylizations for *arbitrary* contents and styles. To achieve these goals, we propose a novel *MicroAST* framework, which will be introduced in detail.

3.1 Overview of MicroAST

As illustrated in Fig. 2, our MicroAST consists of three main components: a micro content encoder E_c, a micro style encoder E_s (coupled with a modulator $\mathcal{M}$), and a micro decoder D. In details, E_c and E_s have the same lightweight architecture which comprises 1 standard stride-1 convolutional (Conv) layer, 2 stride-2 depthwise separable convolutional (DS Conv) layers, and 2 stride-1 residual blocks (ResBlocks). The micro decoder D is mostly symmetrical to the encoders. The modulator $\mathcal{M}$ consists of two subnets as shown in Fig. 4 (see later Sec. 3.2). More detailed architectures can be found in *supplementary material (SM)*[1]. Note that since our model is based on MobileNet (Howard et al. 2017), it can be easily applied to mobile devices. The overall pipeline is as follows:

(1) Extract the main structure of the content image C using the micro content encoder E_c, denoted as $f_c := E_c(C)$.

(2) Extract the style feature of the style image S using the micro style encoder E_s, denoted as $f_s := E_s(S)$.

(3) Convert the style feature f_s into the style modulation signals (a set of learnable parameters) using the modulator $\mathcal{M}$, denoted as $m_s := \mathcal{M}(f_s)$.

(4) Stylize f_c using the micro decoder D, under the dual-modulations (Sec. 3.2) of m_s, *i.e.*, $O := D(f_c, m_s)$.

Training Losses. To achieve style transfer, similar to previous works (Gatys, Ecker, and Bethge 2016; Johnson, Alahi, and Fei-Fei 2016; Huang and Belongie 2017), we leverage a pre-trained VGG-19 (Simonyan and Zisserman 2014) as our loss network to compute the content loss and style loss. We use the perceptual loss (Johnson, Alahi, and Fei-Fei 2016) as our content loss $\mathcal{L}_c$, which is computed at layer $\{relu4\_1\}$ of VGG-19. The style loss $\mathcal{L}_s$ is defined to match the Instance Normalization (IN) statistics (Huang and Belongie 2017), which is computed at layer $\{relu1\_1, relu2\_1, relu3\_1, relu4\_1\}$. *Note that the VGG-19 is only used in our training phase, and our model does not involve any large network at inference.*

To further improve the stylization quality, we also introduce a novel *style signal contrastive (SSC) loss* $\mathcal{L}_{ssc}$ to train our model. This loss could help boost the ability of the micro style encoder to extract more distinct and representative modulation signals for each style (see details in Sec. 3.3).

To summarize, the full objective of our MicroAST is:

$$\mathcal{L}_{full} := \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s + \lambda_{ssc} \mathcal{L}_{ssc}, \qquad (1)$$

where hyper-parameters λ_c, λ_s, and λ_{ssc} define the relative importance of the components in the overall loss function.

3.2 Dual-Modulation

Revisiting Modulation Strategies in AST

(1) AdaIN. (Huang and Belongie 2017) first provided a generic modulation strategy for AST, namely Adaptive Instance Normalization (AdaIN). As illustrated in Fig. 3 (a), AdaIN modulates the content feature f_c with the channel-wise mean $\mu(\cdot)$ and standard deviation $\sigma(\cdot)$ of the style feature f_s.

$$AdaIN(f_c, f_s) := \sigma(f_s)(\frac{f_c - \mu(f_c)}{\sigma(f_c)}) + \mu(f_s). \qquad (2)$$

While AdaIN has obtained great success in recent generative models (Karras, Laine, and Aila 2019; Karras et al. 2020), in style transfer, (Jing et al. 2020) pointed out that there are

[1] https://github.com/EndyWon/MicroAST/releases/download/v1.0.0/MicroAST_SM.pdf

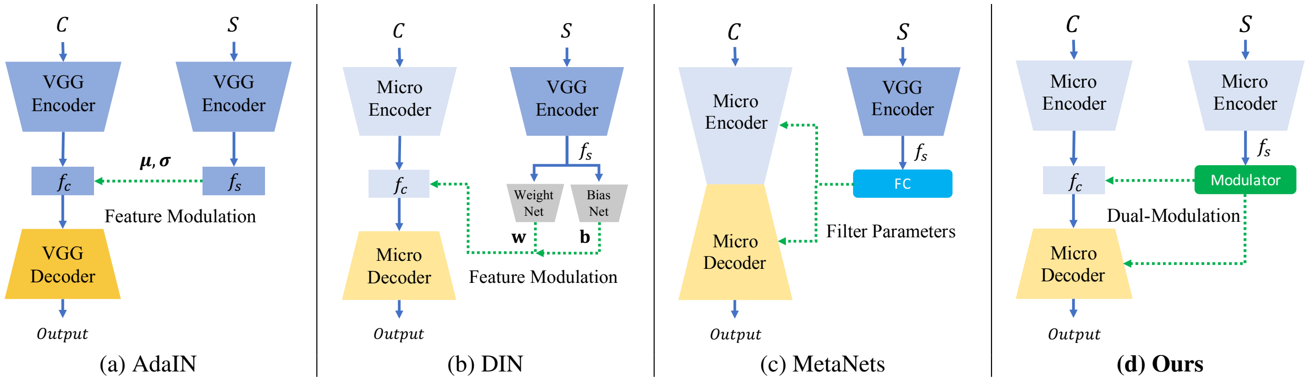


Figure 3: Illustration from modulation to dual-modulation. Existing AST methods either directly modulate/normalize the intermediate features (a, b), or straightforwardly generate the parameters of network filters (c). Our method modulates both intermediate features and network filters (d). "FC" stands for fully connected layers. The details of our dual-modulation are illustrated in Fig. 4.

two critical requirements for AdaIN. (1) The content and style encoders should be identical. (2) The network architecture of the encoders should be complex enough, like VGG. Obviously, requirement (2) is contrary to our task, and requirement (1) hinders the ability of the encoders to extract domain-specific features, especially when they are micro. Therefore, the AdaIN system is not suitable for our task.

(2) DIN. To address the requirement (1) of AdaIN and also train a lightweight network, (Jing et al. 2020) proposed Dynamic Instance Normalization (DIN). As illustrated in Fig. 3 (b), DIN uses a micro content encoder to extract the content feature f_c. The style feature f_s is extracted from a pre-trained sophisticated VGG encoder, along with two subnets to generate the dynamic normalization weight $\mathbf{w}$ and bias $\mathbf{b}$.

$$DIN(f_c, f_s) := \mathbf{w}(\frac{f_c - \mu(f_c)}{\sigma(f_c)}) + \mathbf{b}. \tag{3}$$

These learned dynamic parameters lead to a more accurate alignment of the real complex statistics of style features (Jing et al. 2020), and the micro content encoder and decoder drastically reduce the model size. However, since it still uses the high-cost VGG encoder to extract the style features, the DIN system also cannot be adopted in our task.

(3) MetaNets. Like DIN, (Shen, Yan, and Zeng 2018) also trained a lightweight network for AST. The difference is that they modulate the networks directly instead of the features. As illustrated in Fig. 3 (c), the style image S is first fed into the fixed pre-trained VGG to get the style feature f_s, and then goes through fully connected (FC) layers to construct the filters for each $Conv$ layer in the corresponding image transformation network. The VGG and the FC layers are called "MetaNets". While they can convert an arbitrary new style into a lightweight image transformation network, the high-cost VGG and FC layers lead to the expensive cost of extra memory and slow genuine inference time. Again, the MetaNets system cannot be used for our task.

Dual-Modulation: FeatMod and FilterMod

As analyzed above, the main problem preventing DIN and MetaNets from being applied to our task is the high-cost VGG style encoder. Hence, a simple solution is to replace VGG with a micro encoder to extract the style features. Unfortunately, since the complex pre-trained VGG is also the key of these methods to achieve satisfactory stylizations, the alteration will severely degrade the quality. As shown in the 2^{nd} and 3^{rd} columns of Fig. 5, the VGG style encoder helps DIN and MetaNets to capture complex style patterns like the punctate brushstrokes (top row, best viewed in insets below). However, when replacing VGG with a micro style encoder, these methods consistently learn few texture patterns (bottom row). We attribute it to two main factors: (1) The micro style encoder has limited ability to extract sufficiently complex style features due to the simple network architecture. (2) The style signals injected to guide the stylizations are unitary and inflexible. To address these two problems, we introduce two critical designs in our MicroAST. For the former, we propose a novel *contrastive loss* to boost the ability of the micro style encoder, which will be presented in later Sec. 3.3. For the latter, we propose an innovative *dual-modulation* strategy to inject more sophisticated and flexible style signals to guide the stylizations, which will be introduced in the following.

Our dual-modulation/DualMod strategy seeks to modulate the stylization process from two different dimensions, *i.e.*, intermediate features (feature modulation/FeatMod) and network filters (filter modulation/FilterMod). The motivation comes up from the literature that FeatMod mainly captures the global attributes like rough textures, colors, contrast, and saturation (Huang and Belongie 2017), while FilterMod is particularly good at capturing local changes like different brushstrokes (Alharbi, Smith, and Wonka 2019).

FeatMod. Concretely, our FeatMod adopts the *learned* channel-wise mean and standard deviation as style signals to modulate the intermediate features.

$$\begin{gathered}\boldsymbol{\mu}_s := \mu(f_s), \quad \boldsymbol{\sigma}_s := \sigma(f_s), \\ FeatMod(f_c, (\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s)) := \boldsymbol{\sigma}_s(\frac{f_c - \mu(f_c)}{\sigma(f_c)}) + \boldsymbol{\mu}_s.\end{gathered} \tag{4}$$

Note that it is different from the AdaIN system, as our style

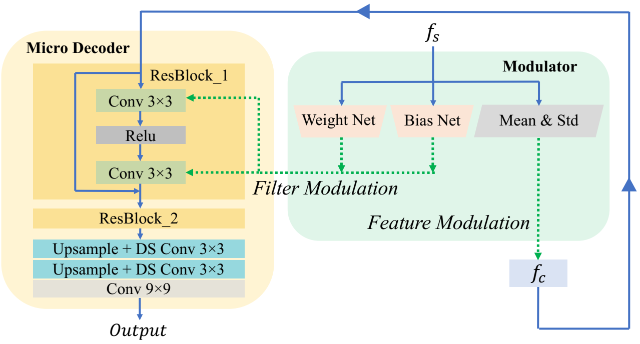


Figure 4: Details of our dual-modulation strategy.

signals are *dynamically* learned from the trainable style encoder E_s, while AdaIN's are *statically* computed from the fixed pre-trained VGG. Also, it is unlike the DIN system, as we use the learned channel-wise mean and standard deviation as style signals, while theirs are computed by two subnets. The reason for using mean and standard deviation is that these statistics can capture the global attributes more effectively (Huang and Belongie 2017), as verified in Fig. 5 (g). By contrast, DIN learns poor on global effects like colors (see Fig. 5 (e)), even with the VGG style encoder (see Fig. 5 (a)).

FilterMod. While FeatMod has been able to capture the global attributes well, it is not enough for style transfer, since the local textures like brushstrokes are also important for artistic styles (Kotovenko et al. 2021). To combat this limitation, we propose a novel FilterMod method in our model.

As illustrated in Fig. 4, the encoded style feature f_s is first converted to the weight $\mathbf{w_s}$ and bias $\mathbf{b_s}$ parameters via two simple subnets (weight net ξ_w and bias net ξ_b). Then, these parameters are injected into decoder D to modulate the $Conv$ filters of the ResBlocks.

$$\begin{aligned} \mathbf{w_s} &:= \xi_w(f_s), \quad \mathbf{b_s} := \xi_b(f_s), \\ FilterMod(D, (\mathbf{w_s}, \mathbf{b_s})) &:= ResBlock(f_c, (\mathbf{w_s}, \mathbf{b_s})) \\ &:= Conv(Relu(Conv(f_c, (\mathbf{w_s}, \mathbf{b_s}))), (\mathbf{w_s}, \mathbf{b_s})) + f_c. \end{aligned} \tag{5}$$

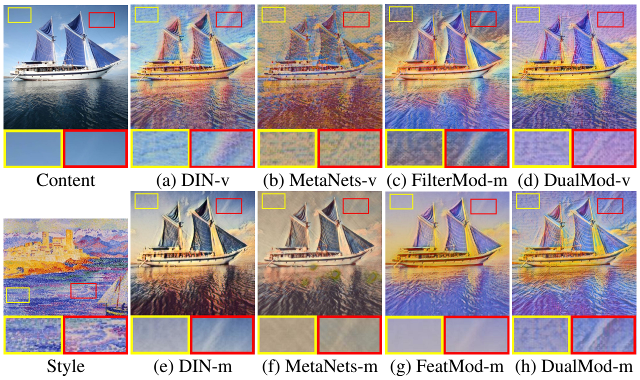
We modulate ResBlocks since they occupy the main complexity of the decoder and dominate the style transfer process (Johnson, Alahi, and Fei-Fei 2016).

Moreover, in many deep learning platforms, it is easier to handle feature maps than filters. Therefore, by the distributive and associative property of convolution, we deduce our FilterMod to an equivalent pseudo FeatMod form as follows:

$$\begin{aligned} Conv(f_c, (\mathbf{w_s}, \mathbf{b_s})) &:= (\mathbf{w_s} * \mathcal{F} + \mathbf{b_s}) \circledast f_c \\ &:= (\mathbf{w_s} * \mathcal{F}) \circledast f_c + \mathbf{b_s} \circledast f_c \\ &:= \mathbf{w_s} * (\mathcal{F} \circledast f_c) + \mathbf{b_s} * f_c, \end{aligned} \tag{6}$$

where $\mathcal{F}$ denotes the convolutional filter, $*$ is the element-wise multiplication with broadcast over the spatial dimensions, and $\circledast$ stands for convolution.

As shown in Fig. 5 (c), under the FilterMod, our MicroAST can capture the challenging punctate brushstrokes well even with the micro style encoder, but the global attributes like colors and contrast are not so good as FeatMod.



Content (a) DIN-v (b) MetaNets-v (c) FilterMod-m (d) DualMod-v

Style (e) DIN-m (f) MetaNets-m (g) FeatMod-m (h) DualMod-m

Figure 5: Comparisons of VGG style encoder (marked by "-v") vs. micro style encoder (marked by "-m"), and modulation vs. dual-modulation (last two columns). *Please see SM for quantitative comparisons and more ablation studies.*

DualMod. Finally, our DualMod is a combination of FeatMod and FilterMod so as to absorb both their merits.

$$\begin{aligned} m_s &:= (\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s, \mathbf{w_s}, \mathbf{b_s}), \\ DualMod(D, f_c, m_s) &:= FeatMod(f_c, (\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s)) + \\ & \quad FilterMod(D, (\mathbf{w_s}, \mathbf{b_s})), \end{aligned} \tag{7}$$

where m_s are the style modulation signals.

As validated in Fig. 5 (h), DualMod helps our MicroAST capture both global attributes and local brushstrokes well. The result is very encouraging and almost on a par with that produced by using VGG style encoder (see Fig. 5 (d)).

3.3 Style Signal Contrastive Learning

As analyzed in Sec. 3.2, due to the simplicity of network architecture, the micro style encoder E_s has limited ability to extract sufficiently complex style representations. Inspired by recent contrastive learning (He et al. 2020; Chen et al. 2020b; Tian, Krishnan, and Isola 2020; Wu et al. 2021) which aims at improving the representative power of neural networks, we propose a novel *style signal contrastive (SSC) loss* $\mathcal{L}_{ssc}$ to boost the style representative ability of E_s.

The core idea of contrastive learning is to pull data points (called "query") close to their "positive" examples, while pushing them apart from other examples that are regarded as "negatives" in the representation space. Therefore, how to construct the "positive" pairs and "negative" pairs is a key problem we need to consider. Intuitively, in our MicroAST, every stylized result is generated under the modulations of a set of specific style signals extracted from a target style image. Therefore, it should possess the similar style signals with the target style image while exhibiting the distinct style signals from other style images. Based on this intuition, given a training mini-batch including N content images $\phi_c = \{C_1, C_2, \ldots, C_N\}$ and N style images $\phi_s = \{S_1, S_2, \ldots, S_N\}$, we first generate N stylized outputs $\phi_o = \{O_1, O_2, \ldots, O_N\}$ (O_i is generated by using C_i as content and S_i as style). Then, for each "query" O_i, we can construct $\phi_p = \{S_i\}$ as its "positive" example, and $\phi_n = \{S_j \in \phi_s | j \neq i\}$ as "negative" examples. Finally,

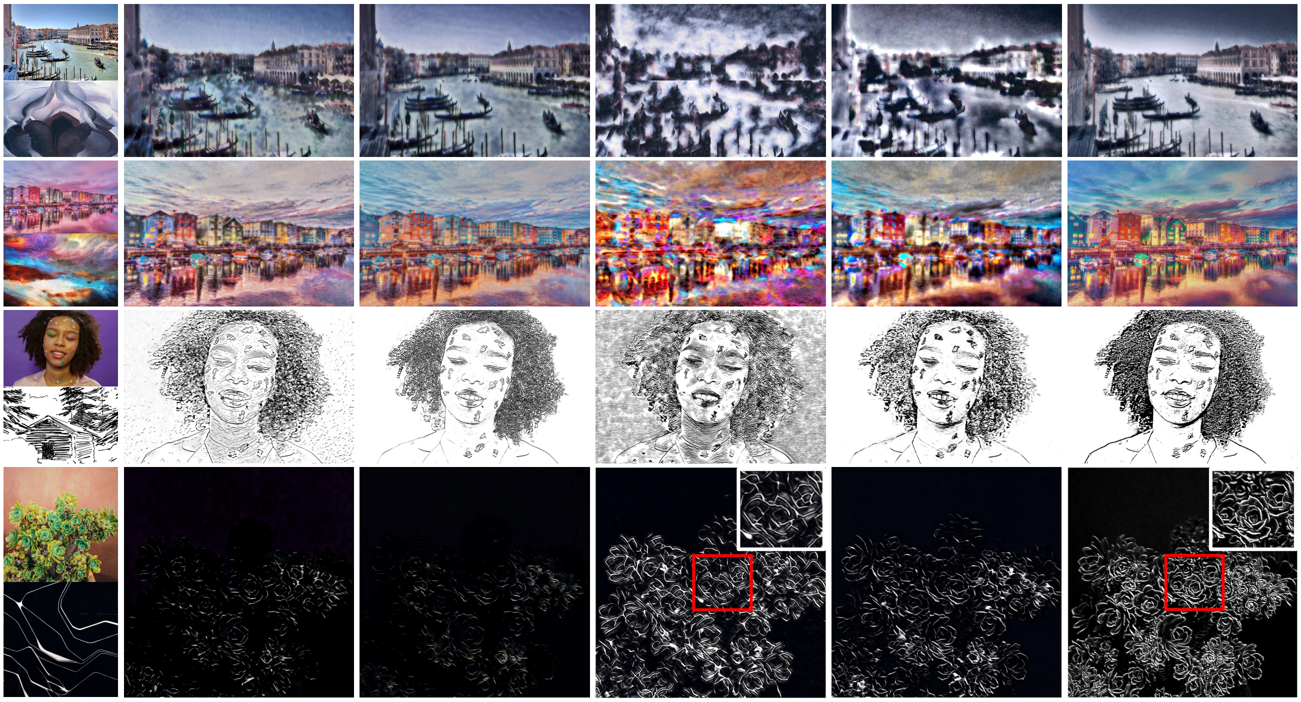


Figure 6: Qualitative comparison with the state of the art on ultra-resolution (4K) images. See more in *SM*.

Method	(a) #Params/10^6	(b) Storage/MB	(c) GFLOPs	(d) Time/sec	(e) SSIM ↑	(f) Style Loss ↓	(g) Preference/%
AdaIN-U	7.011	94.100	5841.9	2.884	0.322	2.401	10.0
LST-U	12.167	48.600	6152.1	2.892	0.441	2.420	24.2
WCT-U	34.239	120.268	19390.1	9.528	0.248	2.226	16.2
Collab-Distill	2.146	9.659	1338.9	7.139	0.292	**2.183**	20.5
MicroAST	**0.472**	**1.857**	**374.9**	**0.522**	**0.531**	2.342	**29.1**

Table 1: Quantitative comparison with the state of the art. Storage is measured in PyTorch model. GFLOPs and Time are measured when the content and style are both 4K images and tested on an NVIDIA RTX 2080 (8GB) GPU. The best results are set in bold. ↑: Higher is better. ↓: Lower is better.

our SSC loss $\mathcal{L}_{ssc}$ is formulated based on the extracted style modulation signals (Eq. (7)) of them.

$$\mathcal{L}_{ssc} := \sum_{i=1}^{N} \frac{\parallel m_{o_i} - m_{s_i} \parallel_2}{\sum_{j\neq i}^{N} \parallel m_{o_i} - m_{s_j} \parallel_2}. \tag{8}$$

$\mathcal{L}_{ssc}$ plays a role of *opposing forces* pulling the style signals of O_i to those of its target style image S_i, and pushing them away from those of other style images. Therefore, it could boost the ability of the micro style encoder to extract more distinct and representative style modulation signals, further improving the stylization quality (see Sec. 4.3).

Discussion. Recently, (Chen et al. 2021a) also introduced contrastive learning for style transfer. There are three main differences: (1) Their contrastive losses are optimized on the *generator* of SANet (Park and Lee 2019) to improve the quality, and their encoders are fixed pre-trained VGG. In contrast, our $\mathcal{L}_{ssc}$ is optimized mainly on the *micro style encoder* to boost its ability in ultra-resolution style transfer. (2) They construct the "positive" pairs and "negative" pairs only within the stylized results, while ours are constructed between the style images and the stylized results. (3) Their contrastive losses are InfoNCE losses (Oord, Li, and Vinyals 2018), which we found is less effective in our task. Thus, we provide a different form of contrastive loss for our $\mathcal{L}_{ssc}$, which is more straightforward and effective in our task.

4 Experimental Results

4.1 Implementation Details

We implement a multi-level DualMod which modulates both the two ResBlocks of the micro decoder D (we omit the modulations for ResBlock_2 in Fig. 2 and Fig. 4 for brevity). The loss weights in Eq. (1) are set to $\lambda_c = 1$, $\lambda_s = 3$, and $\lambda_{ssc} = 3$. We train our MicroAST using MS-COCO (Lin et al. 2014) as content images and WikiArt (Phillips and Mackintosh 2011) as style images. We use the Adam optimizer (Kingma and Ba 2015) with a learning rate of 0.0001 and a mini-batch size of 8 content-style image pairs. During training, all images are loaded with the smaller dimension rescaled to 512 pixels while preserving the aspect ratio, and then randomly cropped to 256×256 pixels for augmentation. Since our MicroAST is fully convolutional, it can handle any input size during testing.

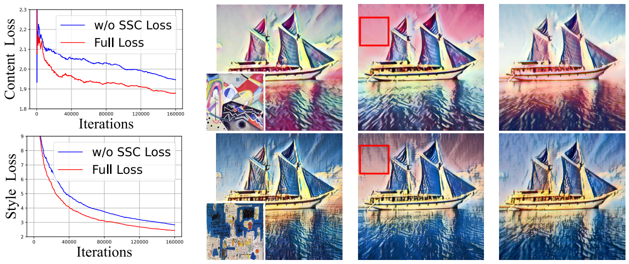
4.2 Comparisons with Prior Arts

We compare our MicroAST with two types of state-of-the-art ultra-resolution AST methods based on (1) model compression (Wang et al. 2020a) and (2) patch-wise stylization (Chen et al. 2022). We directly run the author-released codes with default settings for the compared methods.

Qualitative Comparison. Fig. 6 shows the qualitative comparison results. The patch-wise stylization of URST (Chen et al. 2022) (marked by "-U") can help existing AST methods AdaIN (Huang and Belongie 2017), WCT (Li et al. 2017), and LST (Li et al. 2019) achieve ultra-resolution style transfer. However, AdaIN-U and LST-U often produce less stylized results which retain the colors of content images (*e.g.*, the river color in the 1^{st} row) or transfer insufficient colors of style images (*e.g.*, the 2^{nd} row). WCT-U can transfer more faithful colors, but it often highlights too many textures, resulting in messy stylizations. Built upon WCT, Collab-Distill (Wang et al. 2020a) compresses the VGG-19 models, leading to better-stylized results. Nevertheless, the stylizations are still messy, with spurious boundaries (*e.g.*, the 1^{st} row) and distorted contents (*e.g.*, the 2^{nd} row). In contrast, our MicroAST achieves very promising stylization effects. The contents are sharper and cleaner than WCT-U and Collab-Distill, while the colors and textures are more diverse and adequate than AdaIN-U and LST-U. Moreover, as shown in the bottom two rows[2], it can better preserve the content structures during style transfer, while others either lose the structural details or distort the content structures.

Quantitative Comparison. Tab. 1 shows the quantitative comparison with the state-of-the-art models. We collect 50 ultra-resolution (about 4K) content images and 40 ultra-resolution style images from (Wang et al. 2020a; Chen et al. 2022) and Internet to synthesize 2000 ultra-resolution results, and compute the average Structural Similarity Index (SSIM) (An et al. 2021) and Style Loss (Huang and Belongie 2017) to assess the stylization quality in terms of content preservation and style transformation, respectively. As shown in columns (e) and (f), our method achieves the highest SSIM score and comparable Style Loss, indicating that it can transfer adequate style patterns while better preserving the content affinity. For efficiency (columns (a-d)), our MicroAST is 5-73 times smaller (column (a)) and 6-18 times faster (column (d)) than the state of the art, for the first time enabling *super-fast* AST on 4K ultra-resolution images.

User Study. It is highly subjective to evaluate stylization results. Hence, we conducted a user study for the five approaches. We randomly showed each participant 30 septets of images consisting of the content, style, and five randomly shuffled outputs (AdaIN-U, LST-U, WCT-U, Collab-Distill, and ours). In each septet, they were given unlimited time to select their favorite output in terms of content preservation and stylization effects. We collect 1260 valid votes from 42 subjects and detail the preference percentage of each method in the last column of Tab. 1. The results demonstrate that our stylized images are more appealing than competitors.



(a) Training Curves	(b) Full Model	(c) w/o $\mathcal{L}_{ssc}$	(d) Single-level
SSIM ↑ / Style Loss ↓ :	0.531 / 2.342	0.482 / 2.897	0.554 / 2.686

Figure 7: Ablation study of (a, c) contrastive loss $\mathcal{L}_{ssc}$ and (d) single-level DualMod.

4.3 Ablation Study

With and Without Contrastive Loss $\mathcal{L}_{ssc}$. We demonstrate the effect of our proposed style signal contrastive loss $\mathcal{L}_{ssc}$ in Fig. 7. As shown in column (c), when training our MicroAST without $\mathcal{L}_{ssc}$, the stylization quality is significantly degraded where the colors from one style image may leak into the results stylized by other style images (*e.g.*, the pink color in the red box areas). It indicates that the micro style encoder E_s is floundering in a compromised style representation that may map different styles to similar signals. This problem is alleviated after introducing $\mathcal{L}_{ssc}$ into training, which verifies that contrastive learning indeed helps E_s to learn more distinct and representative style signals. Furthermore, it can also lead to faster and better convergence of content and style optimization and achieve higher SSIM score and lower Style Loss, as validated in column (a) and the bottom row of Fig. 7. *More studies can be found in SM.*

Single-level DualMod vs. Multi-level DualMod. We also compare the results of using single-level DualMod and multi-level DualMod in Fig. 7 (columns (d) and (b)). The multi-level design helps transfer more diverse colors and finer texture details, further improving stylization effects.

5 Conclusion

In this paper, we propose a straightforward and lightweight framework, dubbed MicroAST, for super-fast ultra-resolution arbitrary style transfer. A novel dual-modulation strategy is introduced to inject more sophisticated and flexible style signals to guide the stylizations. In addition, we also propose a new style signal contrastive loss to boost the ability of the style encoder to extract more distinct and representative style signals. Extensive experiments are conducted to demonstrate the effectiveness of our method. Compared to the state of the art, our MicroAST not only produces visually superior results but also is 5-73 times smaller and 6-18 times faster.

[2] To better compare the structures, we perform the same post-processing operations for the images in the bottom two rows.

Acknowledgements

This work was supported in part by the major project of the National Social Science Foundation of China (19ZDA046), the National Key Research and Development Program of China (2020YFC1522704), the National Natural Science Foundation of China (62172365), Zhejiang Elite Program (2022C01222), Key Scientific Research Base for Digital Conservation of Cave Temples of State Administration for Cultural Heritage, and MOE Frontier Science Center for Brain Science & Brain-Machine Integration (Zhejiang University).

References

Alharbi, Y.; Smith, N.; and Wonka, P. 2019. Latent filter scaling for multimodal unsupervised image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1458–1466.

An, J.; Huang, S.; Song, Y.; Dou, D.; Liu, W.; and Luo, J. 2021. ArtFlow: Unbiased Image Style Transfer via Reversible Neural Flows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 862–871.

Champandard, A. J. 2016. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768*.

Chen, H.; Zhao, L.; Qiu, L.; Wang, Z.; Zhang, H.; Xing, W.; and Lu, D. 2020a. Creative and diverse artwork generation using adversarial networks. *IET Computer Vision*, 14(8): 650–657.

Chen, H.; Zhao, L.; Wang, Z.; Zhang, H.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2021a. Artistic Style Transfer with Internal-external Learning and Contrastive Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chen, H.; Zhao, L.; Wang, Z.; Zhang, H.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2021b. DualAST: Dual Style-Learning Networks for Artistic Style Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 872–881.

Chen, H.; Zhao, L.; Zhang, H.; Wang, Z.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2021c. Diverse image style transfer via invertible cross-space mapping. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 14860–14869. IEEE Computer Society.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 1597–1607. PMLR.

Chen, Z.; Wang, W.; Xie, E.; Lu, T.; and Luo, P. 2022. Towards Ultra-Resolution Neural Style Transfer via Thumbnail Instance Normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 393–400.

Cheng, J.; Jaiswal, A.; Wu, Y.; Natarajan, P.; and Natarajan, P. 2021. Style-Aware Normalized Loss for Improving Arbitrary Style Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 134–143.

Chiu, T.-Y. 2019. Understanding generalized whitening and coloring transform for universal style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4452–4460.

Deng, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; and Xu, C. 2021. Arbitrary video style transfer via multi-channel correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, 1210–1217.

Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; and Xu, C. 2022. StyTr2: Image Style Transfer with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11326–11336.

Deng, Y.; Tang, F.; Dong, W.; Sun, W.; Huang, F.; and Xu, C. 2020. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the 28th ACM international conference on multimedia (ACM MM)*, 2719–2727.

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

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 9729–9738.

Hong, K.; Jeon, S.; Yang, H.; Fu, J.; and Byun, H. 2021. Domain-Aware Universal Style Transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 14609–14617.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

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

Jing, Y.; Liu, X.; Ding, Y.; Wang, X.; Ding, E.; Song, M.; and Wen, S. 2020. Dynamic instance normalization for arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 4369–4376.

Jing, Y.; Liu, Y.; Yang, Y.; Feng, Z.; Yu, Y.; Tao, D.; and Song, M. 2018. Stroke controllable fast style transfer with adaptive receptive fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 238–254.

Jing, Y.; Mao, Y.; Yang, Y.; Zhan, Y.; Song, M.; Wang, X.; and Tao, D. 2022. Learning Graph Neural Networks for Image Style Transfer. In *European Conference on Computer Vision (ECCV)*, 111–128. Springer.

Jing, Y.; Yang, Y.; Feng, Z.; Ye, J.; Yu, Y.; and Song, M. 2019. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 26(11): 3365–3385.

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

Kang, M.; and Park, J. 2020. ContraGAN: Contrastive Learning for Conditional Image Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8110–8119.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Kotovenko, D.; Wright, M.; Heimbrecht, A.; and Ommer, B. 2021. Rethinking Style Transfer: From Pixels to Parameterized Brushstrokes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12196–12205.

Li, X.; Liu, S.; Kautz, J.; and Yang, M.-H. 2019. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3809–3817.

Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; and Yang, M.-H. 2017. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 386–396.

Lin, T.; Ma, Z.; Li, F.; He, D.; Li, X.; Ding, E.; Wang, N.; Li, J.; and Gao, X. 2021. Drafting and Revision: Laplacian Pyramid Network for Fast High-Quality Artistic Style Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5141–5150.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 740–755. Springer.

Liu, R.; Ge, Y.; Choi, C. L.; Wang, X.; and Li, H. 2021a. Divco: Diverse conditional image synthesis via contrastive generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 16377–16386.

Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; and Ding, E. 2021b. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6649–6658.

Lu, H.; and Wang, Z. 2022. Universal video style transfer via crystallization, separation, and blending. In *Proc. Int. Joint Conf. on Artif. Intell.(IJCAI)*, volume 36, 4957–4965.

Lu, M.; Zhao, H.; Yao, A.; Chen, Y.; Xu, F.; and Zhang, L. 2019. A Closed-Form Solution to Universal Style Transfer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 5952–5961.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Park, D. Y.; and Lee, K. H. 2019. Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5880–5888.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive learning for unpaired image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 319–345. Springer.

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

Sanakoyeu, A.; Kotovenko, D.; Lang, S.; and Ommer, B. 2018. A style-aware content loss for real-time hd style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 698–714.

Shen, F.; Yan, S.; and Zeng, G. 2018. Neural style transfer via meta networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8061–8069.

Sheng, L.; Lin, Z.; Shao, J.; and Wang, X. 2018. Avatar-Net: Multi-scale Zero-shot Style Transfer by Feature Decoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8242–8250.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 776–794. Springer.

Ulyanov, D.; Lebedev, V.; Vedaldi, A.; and Lempitsky, V. S. 2016. Texture Networks: Feed-forward Synthesis of Textures and Stylized Images. In *International Conference on Machine Learning (ICML)*, 1349–1357.

Wang, H.; Li, Y.; Wang, Y.; Hu, H.; and Yang, M.-H. 2020a. Collaborative Distillation for Ultra-Resolution Universal Style Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1860–1869.

Wang, Z.; Zhang, Z.; Zhao, L.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2022a. AesUST: towards aesthetic-enhanced universal style transfer. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 1095–1106.

Wang, Z.; Zhao, L.; Chen, H.; Li, A.; Zuo, Z.; Xing, W.; and Lu, D. 2022b. Texture reformer: Towards fast and universal interactive texture transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 2624–2632.

Wang, Z.; Zhao, L.; Chen, H.; Qiu, L.; Mo, Q.; Lin, S.; Xing, W.; and Lu, D. 2020b. Diversified Arbitrary Style Transfer via Deep Feature Perturbation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7789–7798.

Wang, Z.; Zhao, L.; Chen, H.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2021. Evaluate and improve the quality of neural style transfer. *Computer Vision and Image Understanding (CVIU)*, 207: 103203.

Wang, Z.; Zhao, L.; Chen, H.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2022c. DivSwapper: towards diversified patch-based arbitrary style transfer. In *Proc. Int. Joint Conf. on Artif. Intell.(IJCAI)*, volume 36, 4980–4987.

Wang, Z.; Zhao, L.; Lin, S.; Mo, Q.; Zhang, H.; Xing, W.; and Lu, D. 2020c. GLStyleNet: exquisite style transfer combining global and local pyramid features. *IET Computer Vision*, 14(8): 575–586.

Wu, H.; Qu, Y.; Lin, S.; Zhou, J.; Qiao, R.; Zhang, Z.; Xie, Y.; and Ma, L. 2021. Contrastive Learning for Compact Single Image Dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10551–10560.

Wu, Z.; Zhu, Z.; Du, J.; and Bai, X. 2022. CCPL: Contrastive Coherence Preserving Loss for Versatile Style Transfer. In *European Conference on Computer Vision (ECCV)*, 189–206. Springer.

Xie, X.; Li, Y.; Huang, H.; Fu, H.; Wang, W.; and Guo, Y. 2022. Artistic Style Discovery With Independent Components. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19870–19879.

Zhang, C.; Zhu, Y.; and Zhu, S.-C. 2019. MetaStyle: Three-Way Trade-off among Speed, Flexibility, and Quality in Neural Style Transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, 1254–1261.

Zhang, Y.; Li, M.; Li, R.; Jia, K.; and Zhang, L. 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.

Zhang, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Lee, T.-Y.; and Xu, C. 2022b. Domain Enhanced Arbitrary Image Style Transfer via Contrastive Learning. *arXiv preprint arXiv:2205.09542*.

Zuo, Z.; Zhao, L.; Lian, S.; Chen, H.; Wang, Z.; Li, A.; Xing, W.; and Lu, D. 2022. Style fader generative adversarial networks for style degree controllable artistic style transfer. In *Proc. Int. Joint Conf. on Artif. Intell.(IJCAI)*, 5002–5009.