AesUST: Towards Aesthetic-Enhanced Universal Style Transfer

Zhizhong Wang*
Zhanjie Zhang*
endywon@zju.edu.cn
Zhejiang University

Lei Zhao†
cszhl@zju.edu.cn
Zhejiang University

Zhiwen Zuo
zzwcs@zju.edu.cn
Zhejiang University

Ailin Li
liailin@zju.edu.cn
Zhejiang University

Wei Xing†
wxing@zju.edu.cn
Zhejiang University

Dongming Lu
ldm@zju.edu.cn
Zhejiang University

ABSTRACT

Recent studies have shown remarkable success in universal style transfer which transfers arbitrary visual styles to content images. However, existing approaches suffer from the aesthetic-unrealistic problem that introduces disharmonious patterns and evident artifacts, making the results easy to spot from real paintings. To address this limitation, we propose AesUST, a novel **Aes**thetic-enhanced **U**niversal **S**tyle **T**ransfer approach that can generate aesthetically more realistic and pleasing results for arbitrary styles. Specifically, our approach introduces an aesthetic discriminator to learn the universal human-delightful aesthetic features from a large corpus of artist-created paintings. Then, the aesthetic features are incorporated to enhance the style transfer process via a novel **Aes**thetic-aware **S**tyle-**A**ttention (AesSA) module. Such an AesSA module enables our AesUST to efficiently and flexibly integrate the style patterns according to the global aesthetic channel distribution of the style image and the local semantic spatial distribution of the content image. Moreover, we also develop a new two-stage transfer training strategy with two aesthetic regularizations to train our model more effectively, further improving stylization performance. Extensive experiments and user studies demonstrate that our approach synthesizes aesthetically more harmonious and realistic results than state of the art, greatly narrowing the disparity with real artist-created paintings. Our code is available at https://github.com/EndyWon/AesUST.

CCS CONCEPTS

• **Applied computing** → **Fine arts**; • **Computing methodologies** → *Rendering*; *Image manipulation*.

KEYWORDS

image style transfer, aesthetic, universal, harmonious, realistic

---

*Both authors contributed equally to this research.
†Corresponding authors.

1 INTRODUCTION

The task of style transfer is to transfer the artistic style from a reference image to a content image, *e.g.*, transferring the style of Vincent Van Gogh's *Sunflower* to an everyday photograph. This task has experienced significant improvements following the seminal work of Gatys *et al.* [14–16], with aspects ranging from efficiency [26, 33, 50], quality [32, 38, 43, 47, 55], generalization [4, 12, 22, 24, 36], diversity [35, 51, 54], and user control [3, 17, 28].

As a central problem of style transfer, universal style transfer (UST) seeks to achieve generalization, quality, and efficiency simultaneously [36]. Depending on the ways of manipulating content and style features, the existing UST approaches can be roughly divided into two categories, *i.e.*, global statistics-based and local patch-based [25]. As representatives of the former, AdaIN [22], WCT [36], LST [34], AAST [21], and ArtFlow [1] transform the content features to match the second-order global statistics of style features. As representatives of the latter, SANet [43], MAST [11], TPFR [49], AdaAttN [40], and IECAST [5] utilize patch-based decorators or attention mechanisms to locally fuse style features into content features. While achieving favorable results, fast speeds, and good generalizations, these approaches often produce disharmonious patterns and evident artifacts, making their results easily distinguishable from real paintings (see Fig. 1). On the contrary, an excellent stylized artwork that can be faked as real should be as realistic as possible in human aesthetics, *i.e.*, its artistic characteristics, such as colors, strokes, tones, textures, *etc*, all exist in harmony and are visually pleasing. Therefore, we define this problem as the aesthetic-unrealistic problem and provide the following analyses:

For global statistics-based approaches [1, 21, 22, 34, 36], the aesthetic-unrealistic problem appears since the global statistics match the disordered style patterns without considering their local distributions [16]. Thus, they may often integrate messy textures into the content target. While for local patch-based approaches [5, 11, 40, 43, 49], the style patterns are integrated to greedily match the local structure distribution of the content image, which is prone to introduce artifacts and disharmonious patterns [58]. Therefore, the leading cause of the aesthetic-unrealistic problem lies in the
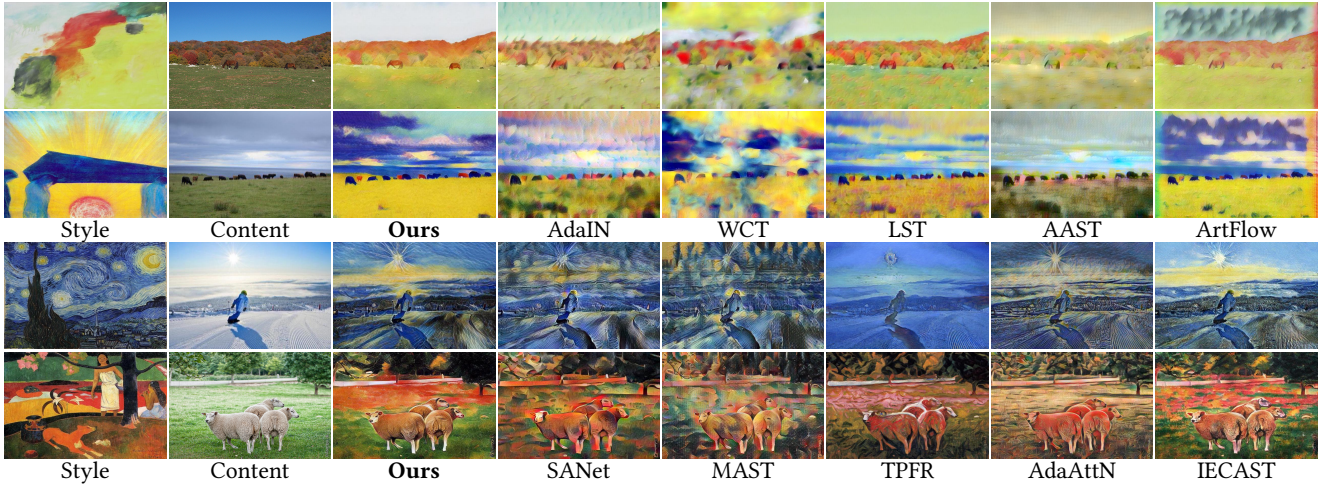
Figure 1: Exemplar stylization results. Existing UST approaches (column 4-8) suffer from the aesthetic-unrealistic problem that introduces disharmonious patterns and evident artifacts, making their results easily distinguishable from real paintings. Our method (column 3) synthesizes aesthetically more realistic and pleasing results, greatly narrowing the disparity with real artist-created paintings.

inappropriate or insufficient considerations on the integration of style patterns.

Based on the above analyses, the current question has been transformed to *how to integrate style patterns in an aesthetically realistic and pleasing way?* However, the subjective characteristics of the problem bring great challenges. There are no absolute standards for measuring the aesthetic quality of a painting [31]. Different artists can have very different ideas towards aesthetics. Therefore, to sidestep such a dilemma, we explore another perspective to address the question. Our insight is based on Immanuel Kant's famous saying, *art is both subjective and **universal***. Each individual person may have his/her own taste, but a given piece of art can also appeal to a large number of people across cultures and history [23]. We argue that the artist-created paintings may share some notable universal characteristics of human-delightful aesthetics (*e.g.*, harmonious patterns and few artifacts), and we can learn and leverage these aesthetics to produce aesthetically more realistic and pleasing results. That is to say, we can utilize these learned universal aesthetics to guide the integration of style patterns, without defining an elaborate aesthetic criterion.

Motivated by the arguments above, we propose a novel ***Aesthetic-enhanced Universal Style Transfer (AesUST)*** approach that aims to generate aesthetically more realistic and pleasing results for arbitrary styles. Our AesUST is inspired by the recent great success of Generative Adversarial Networks (GANs) [18] in generating domain-realistic results. Specifically, we introduce an aesthetic discriminator to learn the universal human-delightful aesthetic characteristics from the domain of a large corpus of artist-created paintings. These paintings are created by numerous artists, thus containing some universal artist-independent and human-delightful aesthetics. The aesthetic discriminator plays two roles here: (i) an opponent that teaches the generator to produce realistic painting-like results via playing a min-max game, (ii) a feature extractor that extracts the aesthetic features to enhance the style transfer process. Notably, we also propose a novel ***Aesthetic-aware Style-Attention***

(AesSA) module to incorporate the aesthetic features with the content and style features. Such an AesSA module enables our approach to efficiently and flexibly integrate the style patterns according to the global aesthetic channel distribution of the style image and the local semantic spatial distribution of the content image, thus helping produce aesthetically more realistic and pleasing results. A remaining issue of our approach is how to perform adversarial training with traditional style transfer training. To this end, we develop a new two-stage transfer training strategy with two aesthetic regularizations to train our model more effectively, further improving stylization performance. We conduct extensive experiments as well as user studies to demonstrate the effectiveness and superiority of our approach. Compared to the state-of-the-art (SOTA) UST algorithms, our AesUST can synthesize aesthetically more harmonious and realistic results, greatly narrowing the disparity with real artist-created paintings.

In summary, our contributions are threefold:

- We reveal the aesthetic-unrealistic problem in existing UST algorithms, and propose a novel *aesthetic-enhanced universal style transfer* approach, *i.e.*, *AesUST*, to generate aesthetically more realistic and pleasing results for arbitrary styles.
- We introduce a novel *aesthetic-aware style-attention (AesSA)* module to achieve global aesthetic-guided and local structure-guided style integration.
- A new two-stage transfer training strategy with two aesthetic regularizations is also developed to train the networks more effectively, further improving stylization performance.

2 RELATED WORK

Global Statistics-based Methods. The seminal work of Gatys *et al.* [14, 16] opened up the era of neural style transfer [25]. They used the global statistics, *i.e.*, the correlations between features extracted from a pre-trained Deep Convolutional Neural Network (DCNN), to represent the style of an image. Albeit the stunning results, their method incurs a rather slow speed in practical usage.
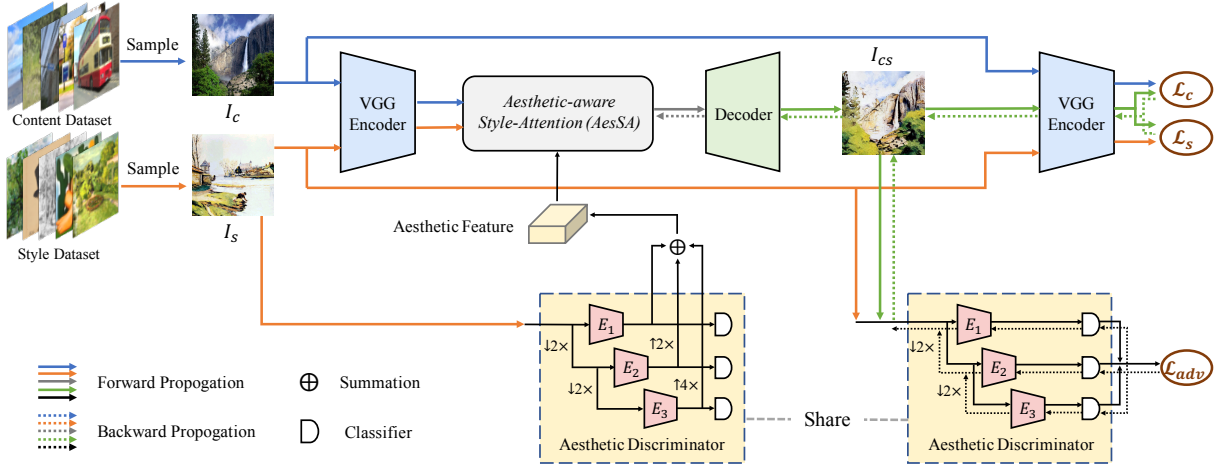
Figure 2: Overview of our proposed AesUST. Note that we apply a two-stage transfer training strategy with a transfer learning fashion: (1) At stage I, the aesthetic discriminator only acts as the discriminator, and we pre-train the generator (AesSA module and decoder) using VGG features only. (2) At stage II, the aesthetic discriminator acts as both the discriminator and feature extractor, and we fine-tune the generator using the aesthetic features extracted from the aesthetic discriminator. The aesthetic discriminator is trained when maximizing the adversarial loss $\mathcal{L}_{adv}$ and fixed when minimizing other losses like content loss $\mathcal{L}_c$ and style loss $\mathcal{L}_s$. For clarity, we omit some stage-specific losses here, i.e., identity loss [43] for stage I, and our proposed aesthetic regularization losses (see Fig. 4) for stage II.

To address this issue, Johnson *et al.* [26] and Ulyanov *et al.* [50] achieved fast style transfer, but the model is limited to a pre-defined style. Thereafter, many successors were advanced to improve the performance in different aspects, including multi-style transfer [4, 12], universal style transfer [10, 22, 24, 34, 36, 41, 42], diversified style transfer [35, 51, 54], unbiased style transfer [1], and domain-aware style transfer [20], *etc.* Unlike these methods, the proposed approach aims to achieve aesthetic-enhanced style transfer, which considers the universal human-delightful aesthetics to improve the quality of universal style transfer.

Local Patch-based Methods. Li and Wand [32, 33] first combined Markov Random Fields (MRFs) and DCNNs to achieve local patch-based neural style transfer. They represented the style patterns by a set of neural patches and greedily integrated them to match the local structure prior of the content image. Later, Chen and Schmidt [9] proposed style-swap for fast patch-based style transfer. Based on it, Sheng *et al.* [47] designed Avatar-Net for zero-shot multi-scale style decoration. Park and Lee [43], Deng *et al.* [11], and Liu *et al.* [40] utilized attention mechanism to integrate style patterns into content features. Other extensions include image analogy [19, 37], semantic style transfer [3, 53, 57], diversified style transfer [56], and multi-modal style transfer [58], *etc.* While these methods can integrate semantically more coherent style patterns into the content features, the greedy structure-guided distributions are prone to introduce artifacts and disharmonious patterns. By contrast, our proposed approach considers both structure-guided and aesthetic-guided distributions to integrate the style patterns, leading to aesthetically more harmonious and realistic results.

Aesthetic-related Methods. While there have been some efforts [2, 31, 55] to explore the automatic evaluation of aesthetic quality, they are confined to the limited aspects, and so far, no unified criterion has been reached. Similar to our method, Sanakoyeu *et*

al. [46], Kotovenko *et al.* [29, 30], Chen *et al.* [6, 7], and Zuo *et al.* [59] also used GAN-based frameworks for style transfer. However, their discriminators are trained to distinguish the aesthetics of a specific artist, *e.g.*, Claude Monet, while ours captures the artist-independent aesthetics. Moreover, their models can only transfer the style of a pre-defined artist, while ours can achieve universal style transfer. Recently, Svoboda *et al.* [49] provided a two-stage peer-regularization model based on graph attention and GAN for arbitrary style transfer, but it suffers from the style-uncontrollable problem [6]. Chen *et al.* [5] used internal-external learning and contrastive learning with GAN to improve the performance of SANet [43], but it may produce implausible stylizations for complex textures and sometimes still introduces noticeable artifacts. Furthermore, Hu *et al.* [21] proposed aesthetic-aware style transfer, but their aesthetics are defined as a combination of only color and texture. By contrast, we provide a more general definition of aesthetics, *i.e.*, the universal human-delightful characteristics learned from a large corpus of artist-created paintings.

3 APPROACH

This part first introduces the overall pipeline of our proposed approach in Sec. 3.1, and then follows it up by providing the details of each component in Sec. 3.2. The two-stage transfer training strategy is specified in Sec. 3.3.

3.1 Overview

Let Φ_c and Φ_s be the datasets of content photographs and real artist-created paintings, respectively. Our goal is to learn a style transformation that can transfer the style of an arbitrary painting $I_s \in \Phi_s$ to a content target $I_c \in \Phi_c$. The key insight is utilizing the aesthetic characteristics learned from Φ_s to enhance the style transformation to synthesize aesthetically more realistic and pleasing

results. To achieve this goal, we propose a novel aesthetic-enhanced universal style transfer framework, termed AesUST.

As shown in Fig. 2, our AesUST consists of four main components: (1) A pre-trained VGG [48] encoder E_{vgg} that projects images into multi-level feature embeddings. (2) An aesthetic-aware style-attention module *AesSA* that appropriately integrates the style patterns into the content features under the guidance of aesthetic and structure distributions. (3) A decoder network D that recovers the feature embeddings to the stylized images. (4) An aesthetic discriminator $\mathcal{D}_a$ that teaches the decoder to generate realistic painting-like results and also extracts the aesthetic features to enhance the style transfer process. The overall pipeline[1] is as follows:

(1) From a content image I_c and style image I_s pair, we first extract the VGG content feature $F_c := E_{vgg}(I_c)$ and style feature $F_s := E_{vgg}(I_s)$ at a certain layer (*e.g.*, *Relu_4_1*) of the encoder E_{vgg}.
(2) We then extract the aesthetic feature of the style image I_s using the aesthetic discriminator $\mathcal{D}_a$, denoted as $F_a := \mathcal{D}_a(I_s)$.
(3) After obtaining the feature maps F_c, F_s, and F_a, we feed them into an *AesSA* module, producing the output feature map $F_{cs} := AesSA(F_c, F_s, F_a)$.
(4) Finally, the stylized output image I_{cs} is synthesized by feeding F_{cs} into the decoder D, *i.e.*, $I_{cs} := D(F_{cs})$.

3.2 Component Details

Encoder-Decoder Module. Following [22]. We employ the pretrained VGG-19 network [48] as our encoder E_{vgg} and fix it all the time. The decoder D is trainable, which mostly mirrors the encoder, but with all pooling layers replaced by the nearest up-sampling.

Aesthetic Discriminator. Inspired by the recent remarkable success of GANs [18] in generating domain-realistic results, we introduce an aesthetic discriminator D_a to learn the universal human-delightful aesthetic features from a large corpus of artist-created paintings. Basically, the aesthetic discriminator D_a is to distinguish between the fake paintings and the real paintings. To achieve so, D_a should perform a sort of encodings of the input paintings before it can judge what paintings are real and what are fake [8, 45]. These encodings, therefore, can be deemed as the notable universal aesthetic features that determine the realism and delicacy of the paintings. Motivated by this, we contend two roles of the aesthetic discriminator: (i) an opponent that teaches the generator to produce realistic painting-like results via playing a min-max game, and (ii) a feature extractor that extracts the aesthetic features to enhance the style transfer process. Such designations bring two main advantages: (1) The framework can be more informative since the adversarial training helps it fuse the domain knowledge of the artist-created paintings. (2) The style transfer networks can be trained more effectively, as the aesthetic discriminator provides additional guidance for better style transfer.

In detail, our aesthetic discriminator follows the multi-scale architecture of [52], which consists of three identical encoders (E_1, E_2, and E_3) and classifiers (C_1, C_2, and C_3). Specifically, we first downsample the real and fake paintings by a factor of 2 and 4 to create an image pyramid of 3 scales. Then, each pair of encoders and classifiers (*e.g.*, E_1 and C_1) is used to differentiate real and fake
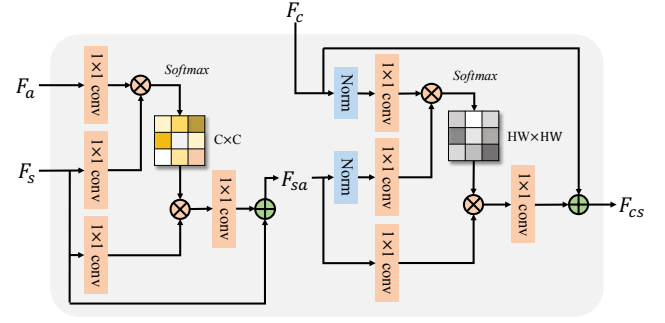
---

[1]The pipeline will vary slightly at different training stages in later Sec. 3.3.

Figure 3: Aesthetic-aware style-attention (AesSA) module. F_c, F_s, and F_a are content, style, and aesthetic features, respectively. "$Norm$" denotes the mean-variance channel-wise normalization.

paintings at one scale. This multi-scale mechanism helps our aesthetic discriminator capture the aesthetic features from different scales, and also guide the style transfer networks to produce more delicate results with coarse-to-fine style patterns. The multi-scale aesthetic discriminator is illustrated in Fig. 2, and the detailed architecture can be found in *appendix*. We sum the encodings of E_1, E_2, and E_3 to obtain the aesthetic features as follows:

$$F_a := \mathcal{D}_a(I_s) := E_1(I_s) \oplus E_2(I_s^{\downarrow^2})^{\uparrow^2} \oplus E_3(I_s^{\downarrow^4})^{\uparrow^4}, \qquad (1)$$

where $\oplus$ denotes element-wise summation, $\downarrow^i$ ($\uparrow^i$) denotes downsampling (up-sampling) operation with factor i.

Aesthetic-aware Style-Attention (AesSA) Module. To incorporate the aesthetic features with the VGG content and style features, we propose the AesSA module, which can adaptively integrate the style patterns into the content features via considering both global aesthetic and local structure features with the attention mechanism. As shown in Fig. 3, AesSA works in two steps: (1) globally enhancing the style features according to the aesthetic channel distributions of the aesthetic features, and (2) locally integrating the enhanced style features according to the semantic spatial distributions of the content features.

Step I: Global aesthetic-guided style enhancement. We first enhance the style features with the aesthetic features extracted from the style images. Inspired by [11, 13, 16] that the channel-wise inner product between the vectorized features can represent the global style well, and the channel attention can effectively improve the feature representation; we introduce the aesthetic attention to globally enhance the VGG style features according to the channel distributions of the aesthetic features. Specifically, given the style feature $F_s \in \mathbb{R}^{C \times H_s \times W_s}$ from VGG and the aesthetic feature $F_a \in \mathbb{R}^{C \times H_s \times W_s}$ from discriminator (where H_s and W_s are the height and width, C is the number of channels), we first obtain their transformed and vectorized features:

$$\begin{aligned}
\hat{F}_a &:= \Gamma(f_a(F_a)) \in \mathbb{R}^{C \times H_s W_s}, \\
\hat{F}_s^1 &:= \Gamma(f_s^1(F_s)) \in \mathbb{R}^{C \times H_s W_s}, \\
\hat{F}_s^2 &:= \Gamma(f_s^2(F_s)) \in \mathbb{R}^{C \times H_s W_s},
\end{aligned} \qquad (2)$$

where Γ denotes the feature vectorization operation, f_a, f_s^1, and f_s^2 are 1×1 learnable convolutions.

Then, we calculate the aesthetic attention between $\hat{F}_a$ and $\hat{F}_s^1$ as follows:

$$A_a := Softmax(\hat{F}_a \otimes (\hat{F}_s^1)^T) \in \mathbb{R}^{C \times C}, \qquad (3)$$

where $\otimes$ is matrix multiplication, T is transpose operation.

Finally, we enhance the style feature F_s through a matrix multiplication and an element-wise summation:

$$F_{sa} := f_{out}^1(\hat{\Gamma}(A_a \otimes \hat{F}_s^2)) \oplus F_s \in \mathbb{R}^{C \times H_s \times W_s}, \qquad (4)$$

where $\hat{\Gamma}$ denotes the reverse operation of Γ that reshapes the vectorized features to the original size, f_{out}^1 is 1×1 learnable convolution.

Step II: Local structure-guided style integration. After enhancing the style features under the aesthetic guidance, we want to integrate them into the content features, so as to achieve aesthetic-aware style transfer. Following SANet [43], we utilize the style attention to locally integrate the enhanced style patterns according to the structure distributions of the content features. Given the aesthetic-enhanced style feature $F_{sa} \in \mathbb{R}^{C \times H_s \times W_s}$ and content feature $F_c \in \mathbb{R}^{C \times H_c \times W_c}$, we first obtain their transformed and vectorized features as follows:

$$\begin{aligned}
\hat{F}_c &:= \Gamma(f_c(Norm(F_c))) \in \mathbb{R}^{C \times H_c W_c}, \\
\hat{F}_{sa}^1 &:= \Gamma(f_{sa}^1(Norm(F_{sa}))) \in \mathbb{R}^{C \times H_s W_s}, \qquad (5) \\
\hat{F}_{sa}^2 &:= \Gamma(f_{sa}^2(F_{sa})) \in \mathbb{R}^{C \times H_s W_s},
\end{aligned}$$

where $Norm$ denotes the mean–variance channel-wise normalization, f_c, f_{sa}^1, and f_{sa}^2 are 1×1 learnable convolutions.

Then, we calculate the style attention between $\hat{F}_c$ and $\hat{F}_{sa}^1$ as follows:

$$A_s := Softmax((\hat{F}_c)^T \otimes \hat{F}_{sa}^1) \in \mathbb{R}^{H_c W_c \times H_s W_s}. \qquad (6)$$

Finally, the output feature F_{cs} is achieved by:

$$F_{cs} := f_{out}^2(\hat{\Gamma}(\hat{F}_{sa}^2 \otimes (A_s)^T)) \oplus F_c \in \mathbb{R}^{C \times H_c \times W_c}, \qquad (7)$$

where f_{out}^2 is 1×1 learnable convolution.

Discussions. In summary, our AesSA performs style integration considering both global aesthetic guidance and local structure guidance. It is closely related to the style attention mechanisms of SANet [43] and MAST [11]. There are, however, two main differences: (1) The inputs of AesSA include not only the VGG content feature F_c and style feature F_s but also the aesthetic feature F_a from the discriminator, integrating different sorts of encoding information. (2) The AesSA is optimized using not only the traditional style transfer training but also the adversarial training with two new aesthetic regularizations. Therefore, our AesSA can appropriately enhance the style patterns and embed them into the content features by mapping the global channel-wise relationship between the style and aesthetic features and the local point-wise relationship between the content and enhanced style features through learning, helping produce aesthetically more realistic and pleasing results.

3.3 Two-Stage Transfer Training

Our framework involves two training processes: adversarial training and traditional style transfer training. The adversarial training is to pursue domain transfer, which encourages that the generated results look like real artist-created paintings. The traditional style transfer training is to achieve universal style transfer, which aims to transfer arbitrary given styles to the content target.

As we introduced in the previous sections, the encodings of the aesthetic discriminator $\mathcal{D}_a$ are utilized to enhance the style transfer process. However, since the ability of the discriminator is gradually improved in the confrontation with the generator, the features it extracts are not always meaningful, especially at the beginning of the adversarial training. The meaningless features will deteriorate the style integration and affect the style transfer training. To overcome this defect, we develop a two-stage transfer training strategy to train the proposed framework in a transfer learning fashion which contains a pre-training stage and a fine-tuning stage. The details of each stage and loss functions are provided below.

Stage I: Pre-training. Since the features extracted from the aesthetic discriminator $\mathcal{D}_a$ are meaningless at the beginning of the adversarial training, we do not inject them into the AesSA module at this stage to guide the style transfer. Instead, we directly exploit the VGG style feature F_s as the aesthetic feature F_a to feed into the AesSA module. In this way, we can pre-train the AesSA module and the decoder to provide a good initialization for the fine-tuning of the next stage. The aesthetic discriminator $\mathcal{D}_a$ here only acts as the discriminator, which plays the min-max game as follows:

$$\begin{aligned}
\max_{\mathcal{D}_a} \min_G \mathcal{L}_{adv}^1 &:= \mathop{\mathbb{E}}_{I_s \sim \Phi_s} [log(\mathcal{D}_a(I_s))] \\
&\quad + \mathop{\mathbb{E}}_{I_c \sim \Phi_c, I_s \sim \Phi_s} [log(1 - \mathcal{D}_a(G(I_c, I_s)))], \qquad (8) \\
G(I_c, I_s) &= D(AesSA(E_{vgg}(I_c), E_{vgg}(I_s), E_{vgg}(I_s))),
\end{aligned}$$

where our generator G consists of a fixed VGG encoder E_{vgg}, a decoder D, and a feature transform module $AesSA$.

For universal style transfer, similar to [22, 43], we use the pre-trained VGG encoder E_{vgg} to compute the content loss $\mathcal{L}_c$ and style loss $\mathcal{L}_s$ as follows:

$$\begin{aligned}
\mathcal{L}_c &:= \sum_{i=1}^{L_c} \parallel Norm(\phi_i(I_{cs})) - Norm(\phi_i(I_c)) \parallel_2, \\
\mathcal{L}_s &:= \sum_{i=1}^{L_s} (\parallel \mu(\phi_i(I_{cs})) - \mu(\phi_i(I_s)) \parallel_2 \qquad (9) \\
&\quad + \parallel \sigma(\phi_i(I_{cs})) - \sigma(\phi_i(I_s)) \parallel_2),
\end{aligned}$$

where μ and σ are the channel-wise mean and standard deviation, respectively. ϕ_i denotes the ith layer of VGG encoder E_{vgg}. We use layers $\{Relu4\_1, Relu5\_1\}$ to compute the content loss, and layers $\{Relu1\_1, Relu2\_1, Relu3\_1, Relu4\_1, Relu5\_1\}$ to compute the style loss.

Moreover, inspired by [11, 43], we also use an identity loss $\mathcal{L}_{id}$ at this stage to constrain the identity mapping relations between content features and style features, which helps better pre-train the generator to maintain the content structure and style characteristics simultaneously [43].

$$\mathcal{L}_{id} := \parallel I_{cc} - I_c \parallel_2 + \parallel I_{ss} - I_s \parallel_2, \qquad (10)$$

where I_{cc} denotes the generated results using a photograph $I_c \in \Phi_c$ as both content and style images, and I_{ss} denotes the generated results using a painting $I_s \in \Phi_s$ as both content and style images.

Full Objective of Stage I. At this stage, the aesthetic discriminator's final objective is

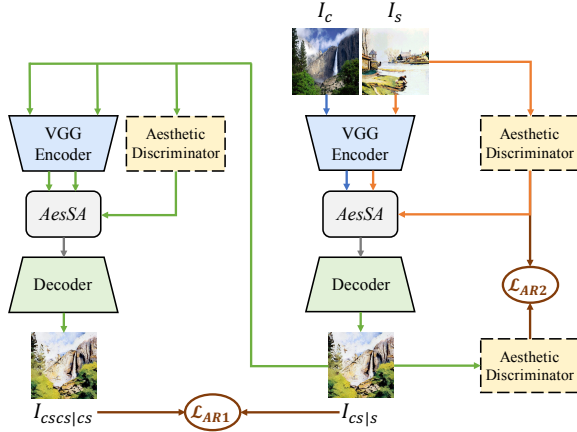$$\max_{\mathcal{D}_a} \mathcal{L}_{adv}^1; \qquad (11)$$

Figure 4: Our proposed two new aesthetic regularizations.

while the generator's final objective is

$$\min_{G} \lambda_1 \mathcal{L}_{adv}^1 + \lambda_2 \mathcal{L}_c + \lambda_3 \mathcal{L}_s + \lambda_4 \mathcal{L}_{id}, \tag{12}$$

where λ_1, λ_2, λ_3, and λ_4 are the trade-off weights which are set to $\lambda_1 = 5$, $\lambda_2 = 1$, $\lambda_3 = 1$, and $\lambda_4 = 50$.

Stage II: Fine-tuning. After rounds of adversarial training at stage I, the aesthetic discriminator D_a is capable of extracting the meaningful aesthetic encodings of input images. Thus, at this stage, we inject the aesthetic feature of style image into the AesSA module to provide the global aesthetic guidance for style transfer. The aesthetic discriminator $\mathcal{D}_a$ here acts as both the discriminator and the feature extractor, which plays a new min-max game as follows:

$$\max_{\mathcal{D}_a} \min_{G} \mathcal{L}_{adv}^2 := \mathbb{E}_{I_s \sim \Phi_s} [log(\mathcal{D}_a(I_s))]$$
$$+ \mathbb{E}_{I_c \sim \Phi_c, I_s \sim \Phi_s} [log(1 - \mathcal{D}_a(G(I_c, I_s)))], \tag{13}$$
$$G(I_c, I_s) = D(AesSA(E_{vgg}(I_c), E_{vgg}(I_s), \mathcal{D}_a(I_s))),$$

here the generator G consists of E_{vgg}, D, $AesSA$, and the aesthetic discriminator D_a. D_a is trained when maximizing $\mathcal{L}_{adv}$, and fixed when minimizing $\mathcal{L}_{adv}$.

We also optimize the content loss $\mathcal{L}_c$ and style loss $\mathcal{L}_s$ defined in Eq. (9) to maintain the generator's generalization ability for universal style transfer. Moreover, to train the networks more effectively and further elevate the stylization performance, we introduce two new aesthetic regularizations at this stage, as shown in Fig. 4.

The form of the first aesthetic regularization $\mathcal{L}_{AR1}$ is similar to the identity loss defined in Eq. (10), but we utilize the generated result as content image, style image, and aesthetic guidance simultaneously.

$$\mathcal{L}_{AR1} := \| I_{cs|s} - I_{cscs|cs} \|_2, \tag{14}$$

where $I_{cs|s}$ denotes the generated result using a photograph $I_c \in \Phi_c$ as content image and a painting $I_s \in \Phi_s$ as both style image and aesthetic guidance. $I_{cscs|cs}$ denotes the generated result using $I_{cs|s}$ as content image, style image, and aesthetic guidance simultaneously. Through this way, we can not only constrain the identity mapping relations like identity loss, but also pull the content feature, style feature, and aesthetic feature extracted from the generated result $I_{cs|s}$ close to those extracted from the content image I_c and style image I_s, thus improving the performance.

To further encourage the guidance of aesthetic features and prevent the networks from ignoring the aesthetic signals, we propose the second aesthetic regularization $\mathcal{L}_{AR2}$ to explicitly enforce the one-to-one mapping between the channel-wise aesthetic features extracted from the style image I_s and the generated result $I_{cs|s}$.

$$\mathcal{L}_{AR2} := \| \mu(\mathcal{D}_a(I_s)) - \mu(\mathcal{D}_a(I_{cs|s})) \|_2$$
$$+ \| \sigma(\mathcal{D}_a(I_s)) - \sigma(\mathcal{D}_a(I_{cs|s})) \|_2 . \tag{15}$$

This kind of aesthetic regularization can help the model preserve aesthetically more pleasing style patterns, which will be clarified by our experiments in later Sec. 4.3.

Full Objective of Stage II. At this stage, the aesthetic discriminator's final objective is

$$\max_{\mathcal{D}_a} \mathcal{L}_{adv}^2; \tag{16}$$

while the generator's final objective is

$$\min_{G} \lambda_5 \mathcal{L}_{adv}^2 + \lambda_6 \mathcal{L}_c + \lambda_7 \mathcal{L}_s + \lambda_8 \mathcal{L}_{AR1} + \lambda_9 \mathcal{L}_{AR2}, \tag{17}$$

where λ_5, λ_6, λ_7, λ_8, and λ_9 are the trade-off weights which are set to $\lambda_5 = 5$, $\lambda_6 = 1$, $\lambda_7 = 1$, $\lambda_8 = 0.5$, and $\lambda_9 = 500$.

4 EXPERIMENTAL RESULTS

4.1 Implementation Details

We follow the multi-level strategy of [43] by integrating two AesSA modules on $Relu4\_1$ and $Relu5\_1$ layers of VGG-19 [48], respectively. The network is trained by using MS-COCO [39] as the content photograph dataset Φ_c, and WikiArt [44] as the artist-created painting dataset Φ_s. Both datasets contain roughly 80000 training images. We use the Adam optimizer [27] with a learning rate of 0.0001 and a mini-batch size of 4 content-style image pairs at both two training stages. Each stage is trained for 80000 iterations. During training, all images are loaded with the smaller dimension rescaled to 512 while preserving the aspect ratio, and then randomly cropped to 256×256 pixels. Since our network is fully convolutional, it can handle arbitrary input size during testing. All experiments are conducted on an NVIDIA RTX 2080 8GB GPU.

4.2 Comparisons

We compare our proposed AesUST against ten SOTA UST approaches including five global statistics-based methods (AdaIN [22], WCT [36], LST [34], AAST [21], and ArtFlow [1]) and five local patch-based methods (SANet [43], MAST [11], TPFR [49], AdaAttN [40], and IECAST [5]).

Qualitative Comparisons. We first present qualitative comparison results in Fig. 5. On the one hand, for global statistics-based methods (top three rows), AdaIN and WCT are prone to produce messy stylized results with conspicuous artifacts and distorted structures. LST, AAST, and ArtFlow can produce cleaner results with better-preserved content structures. However, there are still evident artifacts and disharmonious patterns (*e.g.*, the mixed colors in row 2 and the spurious shadows in row 3). On the other hand, for local patch-based methods (bottom three rows), SANet and MAST often generate distorted contents (*e.g.*, row 4) or structural artifacts (*e.g.*, rows 5-6), making their results unrealistic. TPFR suffers from the style-uncontrollable problem, where the stylized results deviate from the style references. AdaAttN and IECAST perform well in

Figure 5: Qualitative comparisons with SOTA UST approaches. See more in Fig. 1. Zoom-in for better comparison.

Table 1: Quantitative comparisons with SOTA UST approaches.

	WikiArt	**Ours**	AdaIN	WCT	LST	AAST	ArtFlow	SANet	MAST	TPFR	AdaAttN	IECAST
CF	-	**0.420**	0.383	0.346	0.408	0.368	0.391	0.366	0.360	0.401	0.380	0.412
GE + LP	-	1.524	1.480	1.540	1.459	1.461	1.492	**1.552**	1.511	1.220	1.430	1.405
Deception	0.784	**0.568**	0.241	0.172	0.408	0.307	0.356	0.346	0.324	0.414	0.372	0.476
Preference	-	-	0.225	0.183	0.302	0.263	0.350	0.287	0.254	0.247	0.383	0.391
Time (sec.)	-	0.066	0.045	1.163	0.036	1.153	0.382	0.064	0.124	0.582	0.142	0.064

content preservation but cannot transfer complex textures (*e.g.*, rows 4-5) and may produce blended colors (*e.g.*, row 6).

Through considering the universal human-delightful aesthetics, our AesUST synthesizes aesthetically more harmonious and satisfactory results. The transferred textures are more plausible (*e.g.*, the yellow brushes in row 1, the punctate strokes in row 4, and the line abstraction in row 5), the semantic alignment is more accurate (*e.g.*, the sky in row 3), and the style patterns are more harmonious with fewer artifacts (*e.g.*, rows 2 and 6). It shows that AesUST achieves the best performance in stylization effects, significantly narrowing the disparity with real artist-created paintings.

Quantitative Comparisons. We also resort to some quantitative metrics to better evaluate the proposed method.

CF, GE, and LP Scores. Recently, Wang *et al.* [55] proposed three quantifiable factors to evaluate the quality of style transfer, *i.e.*, content fidelity (CF), global effects (GE), and local patterns (LP). In detail, CF measures the faithfulness to content characteristics; GE assesses the stylization quality in terms of the global effects like global colors and holistic textures; LP assesses the stylization quality in terms of the similarity and diversity of the local style patterns. All factors are *the higher the better*. We collect 5 content images and 10 style images to synthesize 50 stylized images for each method and show their average CF, GE, and LP scores in Tab. 1. As marked in bold, our AesUST achieves the highest CF score and the third-highest GE+LP score (slightly lower than SANet [43]
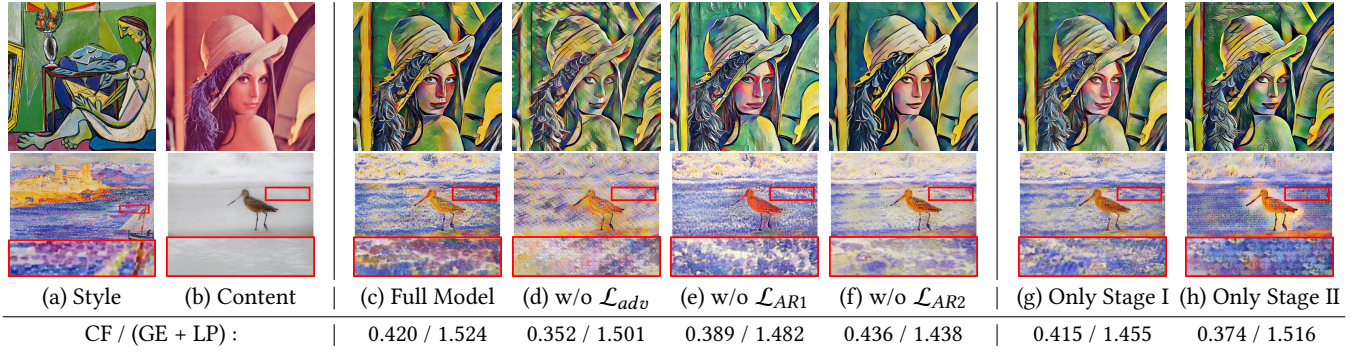
	(a) Style	(b) Content	(c) Full Model	(d) w/o $\mathcal{L}_{adv}$	(e) w/o $\mathcal{L}_{AR1}$	(f) w/o $\mathcal{L}_{AR2}$	(g) Only Stage I	(h) Only Stage II
CF / (GE + LP) :			0.420 / 1.524	0.352 / 1.501	0.389 / 1.482	0.436 / 1.438	0.415 / 1.455	0.374 / 1.516

Figure 6: Ablation study on loss functions (d-f) and training strategy (g-h). Zoom-in for better comparison.

and WCT [36]). It signifies that AesUST achieves the best balance between style transformation and content preservation.

Deception Score. Like [30], to measure the aestheticism and realism of the fake paintings against the real paintings, we conduct a user study to calculate the deception scores as the fraction of times the synthesized images are guessed as "real". We randomly select 20 synthesized images for each method and ask 50 subjects to guess if it is a real painting or not. The results are shown in Tab. 1 ("Deception" row). For comparison, we also report the deception score of the real artist-created paintings from WikiArt [44]. As is evident, the deception score of our AesUST is closest to that of the real artist-created paintings, which further validates the effectiveness of our method.

Preference Score. We conduct A/B Test user studies to compare the stylization effects of our method with the SOTA methods. We randomly select 100 content-style pairs for each subject. In each pair, except for the content and style images, two stylized results generated by our method and a randomly selected SOTA method are displayed in random order. The subjects are asked to choose their preferred outcomes in terms of content preservation and style transformation. We obtain 5000 votes from 50 subjects and show the percentage of votes in Tab. 1 ("Preference" row). The results indicate that our AesUST achieves the best stylization effects.

Efficiency. The bottom row of Tab. 1 shows the run time comparisons on images with a scale of 512×512 pixels. The speed of our AesUST is comparable with SOTA methods such as AdaIN [22], LST [34], SANet [43], and IECAST [5]. Thus, it can practically synthesize stylized images in real time.

4.3 Ablation Study

Loss Analyses. We present ablation study results in Fig. 6 (d-f) to verify the effectiveness of each loss term used for training AesUST. **(1)** Without adversarial loss $\mathcal{L}_{adv}$ (column d), the results exhibit disharmonious patterns and obvious artifacts like the repetitive textures, leading to significant quality degradation. It verifies that adversarial training can help networks synthesize aesthetically more harmonious and realistic results. **(2)** Without the first aesthetic regularization $\mathcal{L}_{AR1}$ (column e), the content preservation and style transformation decline simultaneously. There are some noticeable structural and noisy artifacts, and the style patterns, like the punctate strokes in the bottom row, are not well transferred. **(3)** The second aesthetic regularization $\mathcal{L}_{AR2}$ (column f) can help preserve more pleasing style patterns. Removing it leads to less stylized

results. All above analyses are also supported by the quantitative scores below each column of Fig. 6.

Two-Stage Transfer Training. Fig. 6 (g-h) show the results obtained by each training stage, respectively. **(1)** Training with only stage I (column g) produces less satisfactory results where the colors are not vivid enough (*e.g.*, the top row) and the style patterns are implausible (*e.g.*, the punctate strokes in the bottom row). It suggests that the aesthetic guidance and regularizations of stage II are necessary and effective for producing aesthetically more realistic and pleasing results. **(2)** Directly training with only stage II (column h), yet, also cannot produce satisfactory results. It is because the aesthetic features extracted from the discriminator are meaningless at the beginning of the adversarial training, thus deteriorating the style integration and affecting the style transfer training. Therefore, the pre-training of stage I is also necessary and can force networks to transfer style patterns better while maintaining the content structures.

5 CONCLUSION

In this paper, we reveal the aesthetic-unrealistic problem in the SOTA UST algorithms. Upon analyzing the leading cause of this problem, we propose a novel aesthetic-enhanced UST framework, termed AesUST. Our AesUST introduces an aesthetic discriminator and contends two roles for it, *i.e.*, the discriminator and feature extractor, which learns the universal human-delightful aesthetic features from a large corpus of artist-created paintings. Then, a novel AesSA module is introduced to incorporate the aesthetic features with the content and style features, thus achieving more reasonable and flexible style integration. Furthermore, we also develop a new two-stage transfer training strategy with two aesthetic regularizations to train our model more effectively, further improving stylization performance. Experiments validate that our AesUST can synthesize aesthetically more harmonious and realistic results, greatly narrowing the disparity with real artist-created paintings.

REFERENCES

[1] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. 2021. Art-Flow: Unbiased Image Style Transfer via Reversible Neural Flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 862–871.

[2] Tunç Ozan Aydın, Aljoscha Smolic, and Markus Gross. 2014. Automated aesthetic analysis of photographic images. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 21, 1 (2014), 31–42.

[3] Alex J Champandard. 2016. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768* (2016).

[4] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1897–1906.

[5] Haibo Chen, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, Dongming Lu, et al. 2021. Artistic Style Transfer with Internal-external Learning and Contrastive Learning. *Advances in Neural Information Processing Systems (NeurIPS)* 34 (2021).

[6] Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2021. 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.

[7] Haibo Chen, Lei Zhao, Huiming Zhang, Zhizhong Wang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2021. Diverse image style transfer via invertible cross-space mapping. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 14860–14869.

[8] Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang. 2020. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8168–8177.

[9] Tian Qi Chen and Mark Schmidt. 2016. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337* (2016).

[10] Jiaxin Cheng, Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Prem Natarajan. 2021. Style-Aware Normalized Loss for Improving Arbitrary Style Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 134–143.

[11] 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 28th ACM International Conference on Multimedia*. 2719–2727.

[12] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2017. A learned representation for artistic style. In *International Conference on Learning Representations (ICLR)*.

[13] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3146–3154.

[14] Leon Gatys, Alexander S Ecker, and Matthias Bethge. 2015. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*. 262–270.

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015).

[16] 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.

[17] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3985–3993.

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

[19] Shuyang Gu, Congliang Chen, Jing Liao, and Lu Yuan. 2018. Arbitrary Style Transfer with Deep Feature Reshuffle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8222–8231.

[20] Kibeom Hong, Seogkyu Jeon, Huan Yang, Jianlong Fu, and Hyeran Byun. 2021. Domain-Aware Universal Style Transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[21] Zhiyuan Hu, Jia Jia, Bei Liu, Yaohua Bu, and Jianlong Fu. 2020. Aesthetic-aware image style transfer. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3320–3329.

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

[23] Kiyohito Iigaya, Sanghyun Yi, Iman A Wahle, Koranis Tanwisuth, and John P O'Doherty. 2021. Aesthetic preference for art can be predicted from a mixture of low-and high-level visual features. *Nature Human Behaviour* 5, 6 (2021), 743–755.

[24] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. 2020. Dynamic instance normalization for arbitrary style transfer.

[25] In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4369–4376.

[25] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. 2019. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 26, 11 (2019), 3365–3385.

[26] 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.

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

[28] 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)*. 10051–10060.

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

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

[31] Congcong Li and Tsuhan Chen. 2009. Aesthetic visual quality assessment of paintings. *IEEE Journal of selected topics in Signal Processing* 3, 2 (2009), 236–252.

[32] Chuan Li and Michael Wand. 2016. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2479–2486.

[33] Chuan Li and Michael Wand. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 702–716.

[34] 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)*.

[35] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Diversified texture synthesis with feed-forward networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

[37] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. 2017. Visual attribute transfer through deep image analogy. *ACM Transactions on Graphics (TOG)* (2017).

[38] Tianwei Lin, Zhuoqi Ma, Fu Li, Dongliang He, Xin Li, Errui Ding, Nannan Wang, Jie Li, and Xinbo Gao. 2021. Drafting and Revision: Laplacian Pyramid Network for Fast High-Quality Artistic Style Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5141–5150.

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 740–755.

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

[41] Haofei Lu and Zhizhong Wang. 2022. Universal Video Style Transfer via Crystallization, Separation, and Blending. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*. 4957–4965.

[42] Ming Lu, Hao Zhao, Anbang Yao, Yurong Chen, Feng Xu, and Li Zhang. 2019. A Closed-Form Solution to Universal Style Transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5952–5961.

[43] 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.

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

[45] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

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

[47] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. 2018. Avatar-Net: Multi-scale Zero-shot Style Transfer by Feature Decoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8242–8250.

[48] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[49] 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.

[50] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. 2016. Texture Networks: Feed-forward Synthesis of Textures and Stylized Images.. In *International Conference on Machine Learning (ICML)*. 1349–1357.

[51] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6924–6932.

[52] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8798–8807.

[53] Zhizhong Wang, Lei Zhao, Haibo Chen, Ailin Li, Zhiwen Zuo, Wei Xing, and Dongming Lu. 2022. Texture Reformer: Towards Fast and Universal Interactive Texture Transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2624–2632.

[54] Zhizhong Wang, Lei Zhao, Haibo Chen, Lihong Qiu, Qihang Mo, Sihuan Lin, Wei Xing, and Dongming Lu. 2020. Diversified Arbitrary Style Transfer via Deep Feature Perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7789–7798.

[55] Zhizhong Wang, Lei Zhao, Haibo Chen, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2021. Evaluate and improve the quality of neural style transfer. *Computer Vision and Image Understanding* 207 (2021), 103203.

[56] Zhizhong Wang, Lei Zhao, Haibo Chen, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2022. DivSwapper: Towards Diversified Patch-based Arbitrary Style Transfer. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*. 4980–4987.

[57] Zhizhong Wang, Lei Zhao, Sihuan Lin, Qihang Mo, Huiming Zhang, Wei Xing, and Dongming Lu. 2020. GLStyleNet: exquisite style transfer combining global and local pyramid features. *IET Computer Vision* 14, 8 (2020), 575–586.

[58] Yulun Zhang, Chen Fang, Yilin Wang, Zhaowen Wang, Zhe Lin, Yun Fu, and Jimei Yang. 2019. Multimodal style transfer via graph cuts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5943–5951.

[59] Zhiwen Zuo, Lei Zhao, Shuobin Lian, Haibo Chen, Zhizhong Wang, Ailin Li, Wei Xing, and Dongming Lu. 2022. Style Fader Generative Adversarial Networks for Style Degree Controllable Artistic Style Transfer. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*. 5002–5009.

A DETAILED ARCHITECTURE OF AESTHETIC DISCRIMINATOR

The detailed architecture of our aesthetic discriminator is summarized in Tab. 2, which is adapted from [52].

B RUNTIME CONTROLS

Similar to [22, 43], our AesUST also allows users to control the degree of stylization, interpolate between different styles, transfer styles while preserving colors, and use different styles in different spatial regions. Note that all these controls are only applied at runtime using the same model without re-training.

B.1 Content-style trade-off

We can control the degree of stylization by adjusting the style weights λ_3 in Eq. (12) and λ_7 in Eq. (17) in our main paper during **train** time, or interpolating between feature maps that are fed into the decoder during **test** time. We empirically determine the λ-values in Eq. (12) and Eq. (17) following the previous experience of [22, 43]. Users can adjust them to control the degree of stylization but need to re-train the model, which is inconvenient. In order to achieve runtime control, like [22, 43], we introduce $\alpha \in [0, 1]$ to control the degree of stylization during test time. In detail, given the content image I_c and style image I_s, we first obtain the stylized feature F_{cs} by taking the content image as content input and style image as style input for our model. Then we obtain the feature F_{cc} by taking the content image as both content and style inputs for our model. The α value is used to control the blending of F_{cs} and F_{cc}, obtaining the new stylized feature $F'_{cs} = \alpha F_{cs} + (1 - \alpha)F_{cc}$. The new stylized feature F'_{cs} is fed to the decoder to obtain the final result. In this way, users can control the degree of stylization without re-training the network. The network tries to reconstruct the content image when $\alpha = 0$, and to generate the most stylized image when $\alpha = 1$, as shown in Fig. 7.
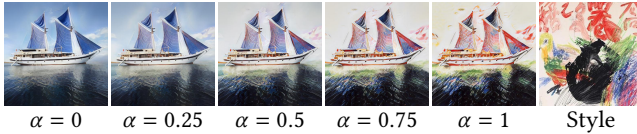

$\alpha = 0$ $\alpha = 0.25$ $\alpha = 0.5$ $\alpha = 0.75$ $\alpha = 1$ Style

Figure 7: Content–style trade-off during runtime.

B.2 Style interpolation

We can interpolate between several style images by blending feature maps F_{cs} transferred from different styles and then feeding the blended feature into the decoder. An example of style interpolation between four different styles is shown in Fig. 8.

B.3 Color and spatial control

Our method can easily achieve color-preserved style transfer similar to [17, 22]. The color distribution of the style image is first manipulated to match that of the content image, and then we perform the style transfer using the color-aligned style image as the style input. Examples are shown in Fig. 9.

Fig. 10 demonstrates that our method can transfer different regions of the content image to different styles. We additionally input

Figure 8: Style interpolation between four different styles.

Style   Content 1   Result 1   Content 2   Result 2

Figure 9: Color control. We show color-preserved style transfer results.

Content   Style   Mask   Result 1   Style   Mask   Result 2

Figure 10: Spatial control examples.

a set of masks to map the spatial correspondence between content regions and styles. A simple mask-out operation is applied to assign the different styles in each spatial region.

C LIMITATIONS AND DISCUSSIONS

In this section, we provide some typical limitations of our method, then analyze the reasons behind them, and finally discuss the possible solutions to address them. Given the simplicity of our proposed framework, we believe there is substantial room for improvement. The further improvement of our approach we leave as future work.

C.1 Mixed textures

As pointed out in our main paper, our method may produce mixed and messy textures for content areas with low feature activations, like the textureless background shown in Fig. 11. It can be attributed to the fact that the textureless areas are hard to recognize (with low feature activations) for a pre-trained VGG network [37]; therefore, the style attention [43] operation in our AesSA module fails to find the correspondence between these areas and those of the style images. This issue may be addressed by incorporating user guidance like [3, 53] or using some texture enhancement operations to highlight the subtle textures in these areas. Moreover, one may

Table 2: Architecture of aesthetic discriminator.

Part	Layer	Kernel Size	Stride	Padding	In Channel	Out Channel	Negative Slope
	Conv	4	2	1	3	64	-
	LeakyReLU	-	-	-	-	-	0.2
	Conv	4	2	1	64	128	-
	InstanceNorm	-	-	-	128	128	-
	LeakyReLU	-	-	-	-	-	0.2
E_1 / E_2 / E_3	Conv	4	2	1	128	256	-
	InstanceNorm	-	-	-	256	256	-
	LeakyReLU	-	-	-	-	-	0.2
	Conv	4	2	1	256	512	-
	InstanceNorm	-	-	-	512	512	-
	LeakyReLU	-	-	-	-	-	0.2
C_1 / C_2 / C_3	Conv	3	1	1	512	1	-
Downsample	Avg Pool	3	2	1	-	-	-
Upsample^2	Nearest	-	1/2	-	-	-	-
Upsample^4	Nearest	-	1/4	-	-	-	-

Content   Style 1   Failure 1   Style 2   Failure 2

Figure 11: Failure cases of type 1. Our method may produce mixed and messy textures for content areas with low feature activations, like the textureless background.

also consider improving the style attention [43] operation in our AesSA module to solve this problem.

C.2 Generalization to out-of-distribution styles

Due to the use of GAN loss, our method may not generalize very well to the style images that excessively deviate from the training distribution, *e.g.*, the simple line pattern shown in Fig. 12. It may be addressed by collecting a more exhaustive training dataset or using some incremental training strategies.

Style   Content 1   Failure 1   Content 2   Failure 2

Figure 12: Failure cases of type 2. Our method may not generalize very well to the style images that excessively deviate from the training distribution, like the simple line pattern.

C.3 Color blending

Our method may produce slight color blending for a few cases. As shown in Fig. 13, the colors of the clothes (marked by red boxes in column 3) seem to be blends of those in the content image and style image. We analyze the reasons may be that the colors in the content image are too vivid compared to those in the style image, thus hard to change via style transfer. The issues can be addressed by removing the colors of the content image as shown in the last column, or training our network with higher style weights λ_3 in Eq. (12) and λ_7 in Eq. (17) in the main paper.

Content   Style   Failure   Remove Color

Figure 13: Failure cases of type 3. Our method may produce slight color blending for a few cases, like the colors of the clothes (marked by red boxes) in the stylized result.