Preserving Structural Consistency in Arbitrary Artist and Artwork Style Transfer

Jingyu Wu[1], **Lefan Hou**[1], **Zejian Li**[2,1*], **Jun Liao**[3], **Li Liu**[3], **Lingyun Sun**[1,4]
[1]Alibaba-Zhejiang University Joint Institute of Frontier Technologies, Zhejiang University, Hangzhou 310027, China
[2]School of Software Technology, Zhejiang University, Ningbo 315048, China
[3]School of Big Data & Software Engineering, Chongqing University, Chongqing 400044, China
[4]Zhejiang-Singapore Innovation and AI Joint Research Lab, Hangzhou 310027, China
{wujingyu, houlefan, zejianlee, sunly, idi}@zju.edu.cn
{liaojun, dcsliuli}@cqu.edu.cn

Abstract

Deep generative models are effective in style transfer. Previous methods learn one or several specific artist-style from a collection of artworks. These methods not only homogenize the artist-style of different artworks of the same artist but also lack generalization for the unseen artists. To solve these challenges, we propose a double-style transferring module (DSTM). It extracts different artist-style and artwork-style from different artworks (even untrained) and preserves the intrinsic diversity between different artworks of the same artist. DSTM swaps the two styles in the adversarial training and encourages realistic image generation given arbitrary style combinations. However, learning style from single artwork can often cause over-adaption to it, resulting in the introduction of structural features of style image. We further propose an edge enhancing module (EEM) which derives edge information from multi-scale and multi-level features to enhance structural consistency. We broadly evaluate our method across six large-scale benchmark datasets. Empirical results show that our method achieves arbitrary artist-style and artwork-style extraction from a single artwork, and effectively avoids introducing the style image's structural features. Our method improves the state-of-the-art deception rate from 58.9% to 67.2% and the average FID from 48.74 to 42.83.

Introduction

Style transfer aims to transfer style from one image to another and preserve structure of target image. Existing methods learn the style from two perspectives. The first perspective is to learn holistic style from a collection of artworks (Liu et al. 2022), but can only produce one kind of stylization, homogenizing the artist-style of different artworks of the same artist (Kotovenko et al. 2019b; Svoboda et al. 2020) and lacking generalization for unseen artists (Chen et al. 2021). While the other is to learn specific style from a single artwork (Wang et al. 2020; Park et al. 2020; Deng et al. 2022; Jing et al. 2022), but might not represent the full scope of an artistic style (Sanakoyeu et al. 2018) and over-adapt to the style image (Cai et al. 2021). Recently, Chen et al. (2021) propose DualAST that learns both holistic artist-style from a collection of artworks and specific artwork-style from a single artwork.

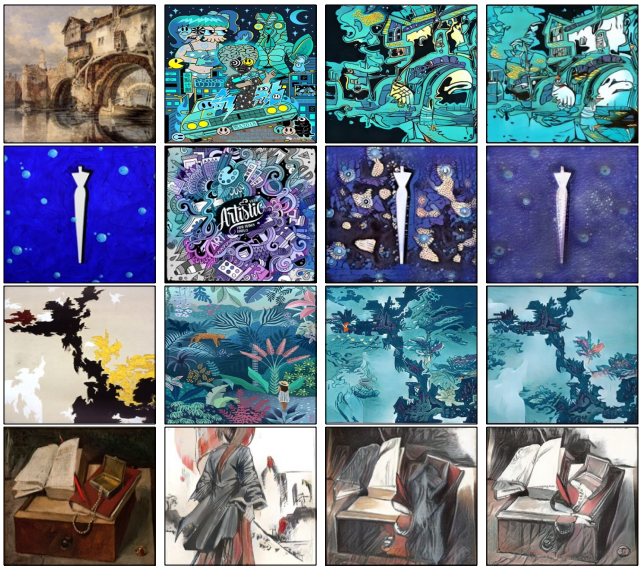

Figure 1: Style transfer results of SwapAE (Park et al. 2020) and ours. From left column to right: we show the content images, style images, generated images using SwapAE, and ours. As shown in the third column, SwapAE mixes the structure of the content image and the style image.

Though DualAST combines two perspectives, it lacks generalization and flexibility. It can only learn one artist-style from a collection of artworks in one network and needs to retrain a new network for another artist (Chen et al. 2021).

To extract both artist-style and artwork-style more effectively from a single artwork, and realize arbitrary artist-style extraction sharing the same network instead of training a new one, we propose a double-style transferring module (DSTM). Our key idea is to decouple the artist-style and artwork-style from the features of an artwork, swap them with those from other images and encourage realistic transferred image generation given arbitrary style combinations. Additionally, we use a classification network to classify latent code of artist-style to avoid trivial solutions and encourage each artist-style to have spatial diversity distance from both holistic and specific perspectives.

*Corresponding author

However, learning style from a single artwork might cause over-adaption to the style image and structural inconsistency of the content image (Cai et al. 2021). Specifically, the generated image introduces some patches that highly resemble the style image, resulting in inferior visual quality (Fig. 1).

Motivated by this, we further propose an edge enhancing module (EEM) to constrain structural consistency in generated images. In detail, EEM gets the edge information from multi-scale and multi-level features learned by holistic image training (Xie and Tu 2015). Compared with the high-frequency domain (Cai et al. 2021) which focuses on the changing of frequency gradient, EEM pays more attention to latent features in the neural network (Fig. 3). Besides, as suggested by Chen et al. (2021), EEM tries to retain only the key structural features rather than all the details in the content image by using a soft edge loss, which is more in line with the goal of style transfer.

Through qualitative and quantitative experiments, our method achieves arbitrary artist-style and artwork-style extraction from a single artwork sharing the same network, and increases the state-of-the-art deception rate by 15%. The artist-style is unique and preserves diversity from both holistic and specific perspectives . Through photo-realistic style transfer, we compare several models using different methods to constrain the structural consistency, such as Whitened (Yoo et al. 2019)(WCT2), swapping code (Park et al. 2020)(SwapAE) and frequency domain (Cai et al. 2021)(FDIT). The results show that our method gets state-of-the-art performance and improves the average FID score from 48.74 to 42.83.

Extensive experiments demonstrate that our approach is highly effective, achieving state-of-the-art performance on six large-scale benchmark datasets. In summary, our main contributions are as follows:

• We propose a double-style transferring module (DSTM) to extract both artist-style and artwork-style from a single artwork, and achieve arbitrary artist-style and artwork-style transfer without training new networks for new artists.

• We introduce an edge enhancing module (EEM) to preserve structural consistency and avoid mixing the structure of content and style images.

• We broadly evaluate our approach across six large-scale datasets with several state-of-the-art style transfer methods. Quantitative and qualitative evaluation results demonstrate the improvements in our method are concise and effective.

Related Work

Neural style transfer. Since Gatys, Ecker, and Bethge (2016) found that the content and style information could be separated from a single artwork by a pre-trained CNN model, multiple works have been introduced in this task. Several recent works try to improve the quality using modern neural networks (Liu et al. 2022; Jing et al. 2022; Deng et al. 2022). Previous works concentrate on extracting style from a single artwork (Huang, Zhang, and Liao 2019; Wang et al. 2020; Jing et al. 2022). They improve the transfer quality from different aspects, such as normalization for generalization (Dumoulin, Shlens, and Kudlur 2016; Huang and Belongie 2017), speed for real-time transferring (Johnson, Alahi, and Fei-Fei 2016; Li and Wand 2016; Ulyanov et al. 2016), structural consistency for image quality (Li et al. 2017; Yoo et al. 2019; Cai et al. 2021) and novel network architecture (Deng et al. 2022; Jing et al. 2022).

Recently, Sanakoyeu et al. (2018) argue that learning style from a single artwork might not represent the full scope of an artistic style and propose AST to learn artist-style from a collection of artworks. Followed by this, many methods (Kotovenko et al. 2019a,b; Svoboda et al. 2020) achieve superior performance, but they still did not achieve satisfying controllable reference-guided stylizations (Chen et al. 2021). Chen et al. (2021) propose DualAST to solve the above challenge by learning holistic artist-style (from a collection of artworks) and specific artwork-style (from a single artwork) simultaneously. However, DualAST only learns one artist-style from a collection of artworks and needs to retrain a new network for different artists. To realize arbitrary artist-style transfer for even unseen artists and ensure the intrinsic diversity of different artworks of the same artist, we propose DSTM. It decomposes the artist-style and artwork-style by swapping features, and uses the classic artwork-style loss and artist classification network to ensure correct extraction.

Generative adversarial networks(GAN).GANs (Goodfellow et al. 2014) have achieved splendid success in many vision tasks, such as image synthesis (Karras et al. 2020, 2021), photo colorization (Kim et al. 2019; Huang et al. 2022) and semantic image synthesis (Qiao, Hancke, and Lau 2022). Recent works try to explore GAN to style transfer. Karras, Laine, and Aila (2019) and Karras et al. (2020) adopt the AdaIN (Huang and Belongie 2017) module for image style transfer. Park et al. (2020) propose an autoencoder architecture that can transfer semantically meaningful structure. Chen et al. (2021) use GAN to identify the artist-style from a collection of artworks in DualAST.

However, existing methods that extract style from a single artwork might over-adapt to it and lose structural consistency. In other words, they can generate images that highly resemble the style image (Cai et al. 2021). AdaConv (Chandran et al. 2021) customizes convolution kernels and SPNST (Cheng et al. 2019) uses edge maps. FDIT constrains the structure features by translating images into high-frequency domain and Fourier domain. In contrast, we propose an edge enhancing module to preserve structural consistency. Different from the high-frequency domain, EEM extracts edge information from VGG layers instead of the frequency gradient. We compare the performance of preserving structural consistency between EEM and high-frequency domain, the results show our method gets better performance.

Method

What is artist-style? Specifically, each artwork of an artist contains his or her own exclusive artist-style but has differences and all his or her artworks show common characteristics. To extract the diverse artist-style in different artwork, we propose a double-style transferring module (DSTM).

We use an autoencoder (Hinton and Salakhutdinov 2006) architecture with three discriminators. We take two artworks

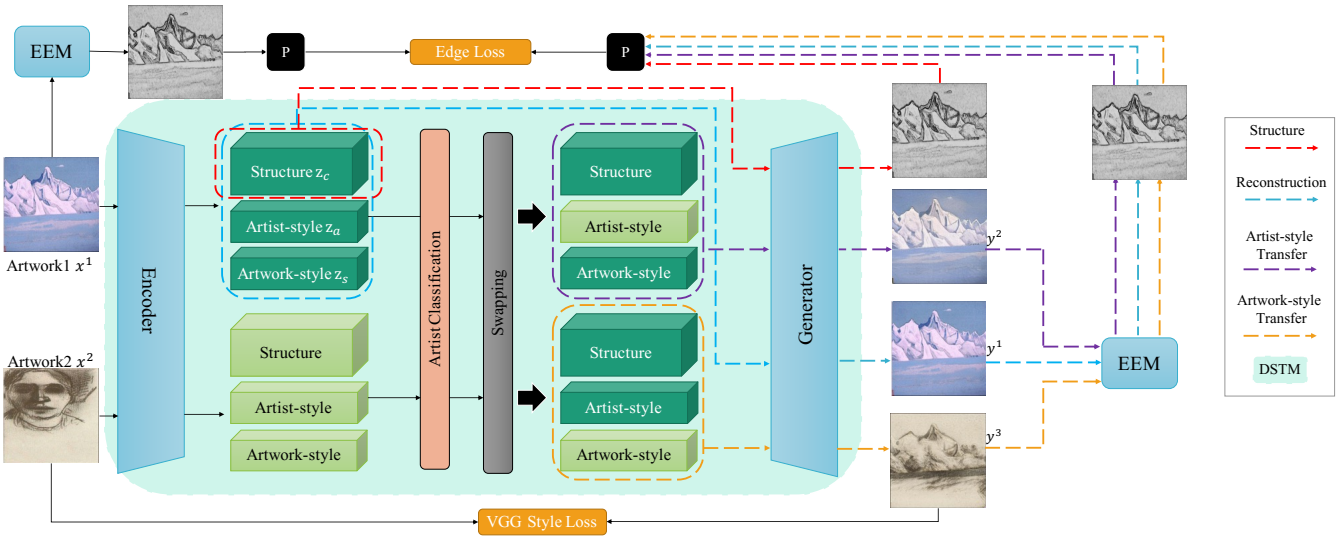

Figure 2: The overview of the proposed double-style transferring module (DSTM) and edge enhancing module (EEM). The key idea of DSTM is to decompose an artwork into structure, artist-style and artwork-style. We add an artist classification network to help decompose the artist-style correctly. DSTM swaps the two styles in the adversarial training and encourages realistic image generation given arbitrary style combinations. EEM gets edge information by deriving multi-scale and multi-level features from the holistic image, then compares the edge information between the content image and output images.

$(\mathbf{x}^1, \mathbf{x}^2)$ as input. The artists of $\mathbf{x}^1$ and $\mathbf{x}^2$ may be the same or different. We aim to learn the artist-style and artwork-style from $\mathbf{x}^2$ and transfer them to an output image using the structure features from $\mathbf{x}^1$.

The structure of our framework (Fig. 2) contains two parts: (1) reconstruction and edge enhancing module (EEM); (2) artist-style and artwork-style transfer (DSTM).

Accurate and Realistic Reconstruction

The Encoder E decomposes the input images $\mathbf{x}^1, \mathbf{x}^2 \sim \mathbf{X} \subset \mathbb{R}^{H \times W \times 3}$ into latent space $\mathbf{Z}$. The core objective of the reconstruction part is to reconstruct an artwork with high quality and accuracy using a reconstruction loss:

$$\mathcal{L}_{rec}(E, G) = \mathbb{E}_{\mathbf{x^1} \sim \mathbf{X}}[||\mathbf{x^1} - G(E(\mathbf{x^1}))||_1] \quad (1)$$

To improve the quality of the reconstruction image and make them more realistic, we use a discriminator D. The adversarial loss (Goodfellow et al. 2014) for generator G and encoder E is calculated as:

$$\mathcal{L}_{GAN,rec}(E, G, D) = \mathbb{E}_{\mathbf{x^1} \sim \mathbf{X}}[-\log(D(G(E(\mathbf{x^1}))))] \quad (2)$$

Decomposable Latent Codes and Double-style Transferring Module

To achieve arbitrary artist and artwork transfer, we first divide the latent space $\mathbf{Z}$ into three components $\mathbf{z} = (\mathbf{z}_c, \mathbf{z}_s, \mathbf{z}_a)$. We swap the components with those from other images and still produce realistic images with GAN Loss:

$$\begin{aligned}&\mathcal{L}_{GAN,swap}(E, G, D) = \mathbb{E}_{\mathbf{x}^1, \mathbf{x}^2 \sim \mathbf{X}, \mathbf{x}^1 \neq \mathbf{x}^2} \\ &[-(\log(D(G(\mathbf{z}_c^1, \mathbf{z}_s^2, \mathbf{z}_a^1))) + \log(D(G(\mathbf{z}_c^1, \mathbf{z}_s^1, \mathbf{z}_a^2))))]\end{aligned} \quad (3)$$

where $\mathbf{z}_c^1, \mathbf{z}_s^1, \mathbf{z}_a^1$ and $\mathbf{z}_c^2, \mathbf{z}_s^2, \mathbf{z}_a^2$ are structure, artwork-style and artist-style components of $E(\mathbf{x}^1), E(\mathbf{x}^2)$.

As for the artwork-style $\mathbf{z}_s$ extraction control, similar to the previous works (Park and Lee 2019; Chen et al. 2021), we leverage a fixed pre-trained VGG-19 network ϕ to compute. We formulate the artwork-style loss $\mathcal{L}_s$ as:

$$\begin{aligned}\mathcal{L}_s(E, G) = &\sum_{i=1}^{n} ||\mu(\phi_i(G(\mathbf{z}_c^1, \mathbf{z}_s^2, \mathbf{z}_a^1))) - \mu(\phi_i(\mathbf{x}^2))||_2 \\ &+ \sum_{i=1}^{n} ||\sigma(\phi_i(G(\mathbf{z}_c^1, \mathbf{z}_s^2, \mathbf{z}_a^1))) - \sigma(\phi_i(\mathbf{x}^2))||_2\end{aligned} \quad (4)$$

Where μ and σ are channel-wise mean and standard deviation, respectively. ϕ_i denotes a layer in VGG-19 used to compute the artwork loss.

We also use a patch co-occurrence discriminator D_{style} to improve artwork-style transfer performance on small patches (Park et al. 2020). The loss function of D_{style} is:

$$\begin{aligned}&\mathcal{L}_{Patch}(E, G, D_{style}) = \mathbb{E}_{\mathbf{x}^1, \mathbf{x}^2 \sim \mathbf{X}, \mathbf{x}^1 \neq \mathbf{x}^2} \\ &[-\log(D_{style}(crop(G(\mathbf{z}_c^1, \mathbf{z}_s^2, \mathbf{z}_a^1)), crop(\mathbf{x}^2)))]\end{aligned} \quad (5)$$

where $crop$ selects a random patch of size $1/8$ to $1/4$ of the full image dimension on each side.

To ensure that the artist-style $\mathbf{z}_a$ decomposes correctly and to avoid trivial solutions, a classification network is used to classify the artist-style. We add an average pooling layer before classification to prevent E homogenizing the artist-style of different artworks of the same artist and preserve diversity. The loss function of the classification network is:

$$\mathcal{L}_{CLS,artist}(E) = \mathcal{L}_{ce}(Cls(P(\mathbf{z}_a)), l_{artist}) \quad (6)$$

where Cls, P and l represent the classification network, average pooling layer and label, respectively. $\mathcal{L}_{ce}$ represents a

cross-entropy loss.

Furthermore, we swap latent components $\mathbf{z}_a^1, \mathbf{z}_a^2$ and generate artist-style transferred image $G(\mathbf{z}_c^1, \mathbf{z}_s^1, \mathbf{z}_a^2)$. An artist discriminator D_{artist} ensures that the decoupled artist-style is used in the generator and the artist-style transferred image has the correct artist-style. The loss function of D_{artist} is:

$$\mathcal{L}_{Artist}(E, G, D_{artist}) = \mathbb{E}_{\mathbf{x}^1, \mathbf{x}^2 \sim \mathbf{X}, \mathbf{x}^1 \neq \mathbf{x}^2} [-\log(D_{artist}(G(\mathbf{z}_c^1, \mathbf{z}_s^1, \mathbf{z}_a^2)))] \tag{7}$$

The total double-style transferring module loss is:

$$\mathcal{L}_{DSTM} = 0.5\mathcal{L}_{GAN,swap} + \mathcal{L}_s + \mathcal{L}_{Patch} + \mathcal{L}_{Artist} + 0.5\mathcal{L}_{CLS,artist} \tag{8}$$

Edge Enhancing Module

Since learning style from a single artwork might over-adapt to it and mix the structure of the content image and style image (Cai et al. 2021) in Fig. 1 (3rd col), we further propose an Edge Enhancing Module (EEM) to preserve structural consistency between content and transferred images. The structure of EEM is inspired by HED (Xie and Tu 2015), which can learn the edge features from multi-scale and multi-level, and do the holistic prediction. EEM aims to get the edge information from the input content image and three output images (reconstruction $\mathrm{y}^1 = G(\mathbf{z}_c^1, \mathbf{z}_s^1, \mathbf{z}_a^1)$, artwork-style transfer $\mathrm{y}^2 = G(\mathbf{z}_c^1, \mathbf{z}_s^2, \mathbf{z}_a^1)$ and artist-style $\mathrm{y}^3 = G(\mathbf{z}_c^1, \mathbf{z}_s^1, \mathbf{z}_a^2)$). Then, EEM reduces the distance of edge information between the content image and output images by enforcing a soft edge loss $\mathcal{L}_{eg_{soft}}$ ($\mathcal{L}_{eg1}$ and $\mathcal{L}_{eg2}$):

$$\mathcal{L}_{eg1}(E, G) = \mathbb{E}_{\mathbf{x}^1, \mathbf{x}^2 \sim \mathbf{X}, \mathbf{x}^1 \neq \mathbf{x}^2} \frac{1}{n} \sum_{i=1}^{n} ||P(EEM(\mathrm{y}^i)) - P(EEM(\mathbf{x}^1))||_2^2 \tag{9}$$

where P is an average pooling layer that tries to capture the key information from the output of EEM. Compared with typical loss, the soft edge loss is more in line with the goal of style transfer: preserving key structural information rather than preserving all structural details (Chen et al. 2021).

In addition, to empower the encoder E and generator G with the ability to capture and perceive latent structural features, respectively, we take only the structural component $\mathbf{z}_c$ as the input of G and calculate the soft edge image loss between the generated image and the edge of content image. We demonstrate its effectiveness in ablation studies (Tab. 3):

$$\mathcal{L}_{eg2}(E, G) = \mathbb{E}_{\mathbf{x}^1 \sim \mathbf{X}} ||P(G(\mathbf{z}_c^1)) - P(EEM(\mathbf{x}^1))||_2^2] \tag{10}$$

Notice that recent work (Cai et al. 2021) uses the high-frequency domain to constrain the structural consistency. Though the high-frequency domain is similar to edge detection, it pays more attention to the edge information reflected by the change of frequency gradient and does not consider the actual object. We demonstrate the difference between high-frequency and edge detection in Fig. 3. When edge color is close to the background color, the edges cannot be detected clearly in the high-frequency domain.

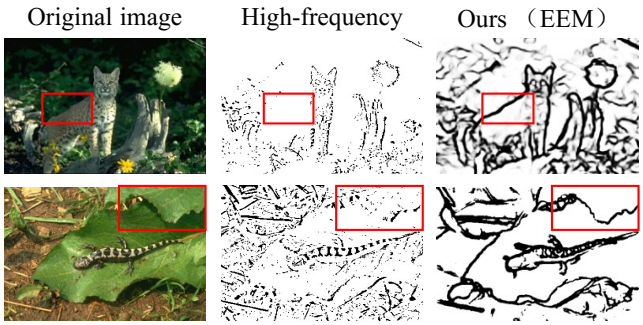

Figure 3: Comparisons of the performance between high frequency domain (kernel size = 21 (Cai et al. 2021)) and our method from BSDS 500 (Arbelaez et al. 2010). For edge color that is close to the background color, the edges cannot be detected clearly in the high-frequency domain.

Overall Training

Our final objective function for the encoder and generator is $\mathcal{L}_{total} = \mathcal{L}_{rec} + 0.5\mathcal{L}_{GAN,rec} + \mathcal{L}_{DSTM} + \mathcal{L}_{eg_{soft}}$. The detailed implementations of our networks are in the appendix.

Experiments

We conduct extensive experiments and comparisons to evaluate our method. First, we show qualitative comparisons of artwork-style, artist-style and photo-realistic style transfer results generated by our model and other baselines. Next, quantitative results are presented. Finally, we conduct ablation studies to validate the effectiveness of each component.

Experimental Setup

Datasets. For artwork-style and artist-style transfer, we use WikiArt (Karayev et al. 2013) for content and style images. For photo-realistic style transfer, we evaluate the performance of preserving structural consistency on the following five large datasets: (1) LSUN Church (Yu et al. 2015), (2) LSUN Bedrroms (Yu et al. 2015), (3) Flickr Faces HQ (FFHQ) (Karras, Laine, and Aila 2019), (4) Flickr Waterfalls (100k self-collected images) (Cai et al. 2021), (5) CelebA-HQ (Karras et al. 2017). All datasets are at a resolution of 256px in training. Note that in the inference stage, both the content and style images can be of any size due to the fully convolutional architecture.

Baselines. For artwork-style and artist-style transfer, we use Gatys, Ecker, and Bethge (2016), AdaIN (Huang and Belongie 2017), WCT (Li et al. 2017), Svoboda et al. (2020), SwapAE (Park et al. 2020) and DualAST (Chen et al. 2021) as our baselines. Among them Gatys, Ecker, and Bethge (2016), AdaIN (Huang and Belongie 2017), WCT (Li et al. 2017), SwapAE (Park et al. 2020) learn artwork-style from a single artwork, while DualAST (Chen et al. 2021) and Svoboda et al. (2020) learn artist-style from a collection of artworks. To validate the performance of preserving structural consistency on photo-realistic style transfer, we evaluate our method by comparing with three

	Gatys et al.	AdaIN	WCT	Svoboda et al.	SwapAE	DualAST	Ours
Deception Rate (↑)	0.206	0.065	0.027	0.278	0.336	0.589	**0.672**
Visual Quality(%) (↑)	0.088	0.068	0.018	0.112	0.188	0.213	**0.313**
Style Controllability(%) (↑)	0.159	0.112	0.134	0.098	0.089	0.191	**0.217**

Table 1: The deception rate, visual quality and style controllability for different methods. The best scores are reported in bold.

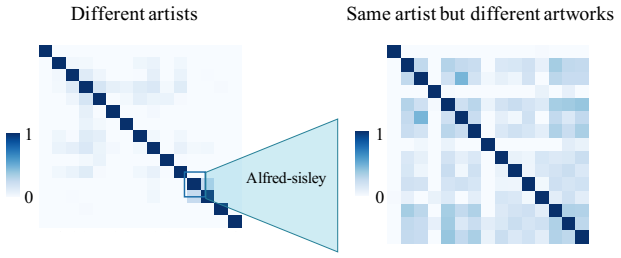

Figure 4: Diverse distances of artist-style from different artists and same artist but different artworks. All artists and artworks are from test set. The artist-style of unseen artworks are diverse from holistic and specific perspectives.

state-of-the-art architectures FDIT (Cai et al. 2021), SwapAE (Park et al. 2020), STROTSS (Kolkin, Salavon, and Shakhnarovich 2019), and three GAN inversion models StyleGAN3 (Karras et al. 2021), StyleGAN2 (Karras et al. 2020), Im2StyleGAN (Anokhin et al. 2020).

Qualitative Results

To validate the superiority of our method, we compare our artist-style and artwork-style transfer results with those of the aforementioned six baselines in Fig. 5. Gatys, Ecker, and Bethge (2016) may encounter the incorrect local minimum (rows 2, 5 and 6). AdaIN sometimes introduces the colors or patterns that do not exist in the style images (rows 1, 4 and 5). WCT struggles in the face of complicated internal lines in the content images (rows 2, 4 and 5). Svoboda et al. (2020) only learn the holistic artist-style from the whole artwork dataset, resulting in uncontrollable stylizations and introducing the holistic color even do not exist in the style image (row 2 and 5). DualAST learns the features of artist-style but it needs to retrain different networks for different artists. When the content images with complex structures or similar overall color, it cannot learn structural features correctly (rows 1, 5 and 6).

For the artistic-style transfer, take the first row as an example, we transfer artist-style of Cézanne to sketch northern renaissance (artwork-style) artwork. The outline of transferred image is thicker and the image is more stereoscopic which are the characteristics of Cézanne: the outline of an object is often drawn with thick black lines to make the artwork deeper and more textured. More details are in the appendix.

In Fig. 7, we show photo-realistic style transfer results to validate the performance of preserving structural consistency on Church, Bedroom, Waterfalls and FFHQ, compared with StyleGAN3, StyleGAN2, STROTSS, SwapAE and FDIT. StyleGAN2 generates fuzzy images and loses the structural details (rows 1, 2 and 3). STROTSS introduces the noise or albefaction in the corner of images (rows 1 and 3). SwapAE introduces the structural features of the style images (row 3). FDIT constrains structure by high-frequency and Fast Fourier Transform but facing edge color close to the background, the performance decreases (row 4).

To prove the generalization, we also provide diverse distances between different *unseen artists* and the same artist but different artworks (Fig. 4). The artist-style is diverse and has spatial distance from holistic and specific perspectives.

Quantitative Results

In this section, we use three quantitative evaluation metrics for comparison: deception rate, visual quality, and style controllability. We also use FID, LPIPS and user study to test the performance of constraining structural features on photo-realistic style transfer. Thus, a total of six evaluation metrics are used to better evaluate our method in different aspects.

Deception rate. Sanakoyeu et al. (2018) propose it to quantitatively show the performance of artist-style transfer. The main idea is to train a VGG-16 to classify which artist each artwork belongs to. Then the VGG-16 predicts which artist the transferred image belongs to. Finally, we calculate the fraction of times that the network predicts correctly as the deception rate. We test our method and six baselines in Tab. 1. Our method outperforms other methods and achieves the highest score, improving it from 58.9% to 67.2%.

User study. Due to the style transfer task being a highly subjective task, user study is widely adopted in the previous works (Park et al. 2020; Chen et al. 2021; Liu et al. 2022). Here we use two user evaluation metrics: visual quality and style controllability to evaluate the user preference of each method. We select 20 content and style image pairs as the input of the above several compared methods, yielding 20 transferred images for each method. We randomly ordered all the transferred images and show them to participants and ask them two questions. The first is which image can best represent the style of the target image while ensuring the quality of the generated image (*Visual Quality*). The other is which image learns the most characteristics from the style image (*Style Controllability*). We collect 1000 votes from 50 participants for each question and report the result in the second row of Tab. 1, where we can see the results obtained by our method are more popular than those of other methods.

FID. FID (Heusel et al. 2017) measures the distance between the ground truth image and the feature vector of the generated image using the InceptionV3 (Szegedy et al. 2016) network. We adopt the FID as the measure of image quality, the smaller values the better. Our method improves

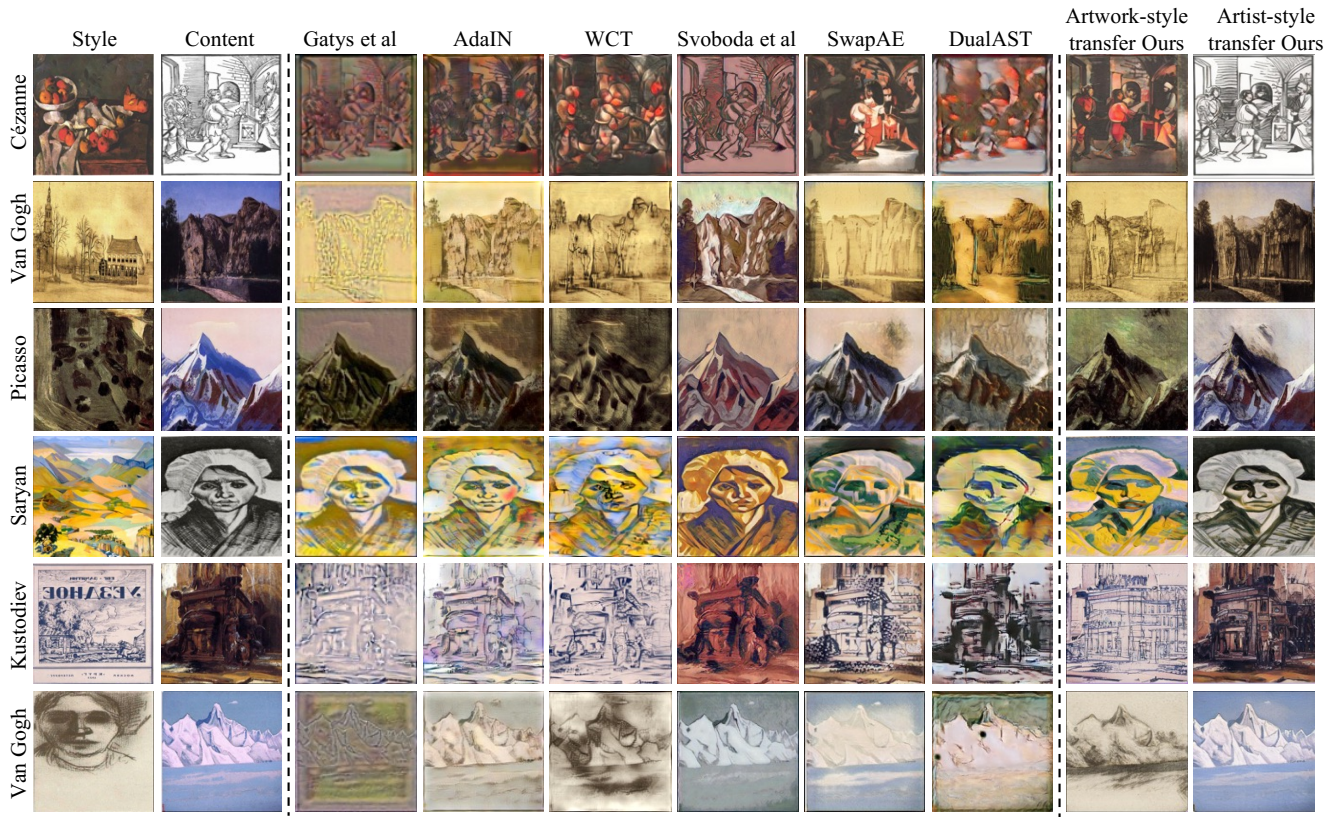

Figure 5: Qualitative comparisons. The first column shows the content images from different artists. The second column shows the style images. The rest columns show the stylization results generated by different baseline models.

	FID (↓)					LPIPS Reconstruction (↓)				
Model	Church	Waterfalls	FFHQ	CelebA-HQ	Average	Church	Waterfalls	FFHQ	CelebA-HQ	Average
Im2StyleGAN	219.5	267.25	123.13	70.14	170.01	**0.186**	0.281	0.174	0.185	0.207
StyleGAN2	57.54	57.46	81.44	56.69	63.28	0.377	0.384	0.215	0.175	0.288
StyleGAN3	55.35	51.25	63.17	42.64	53.10	0.233	0.274	0.140	0.161	0.202
Swap AE	52.34	50.90	59.83	43.47	51.64	0.227	0.238	**0.074**	0.073	0.153
FDIT	48.21	48.76	**55.96**	42.02	48.74	0.205	0.242	0.075	0.076	0.140
Ours	**42.71**	**34.52**	56.98	**37.10**	**42.83**	0.192	**0.223**	0.078	**0.073**	**0.142**

Table 2: Comparison of FID and LPIPS score on four diverse datasets. The best scores are reported in bold.

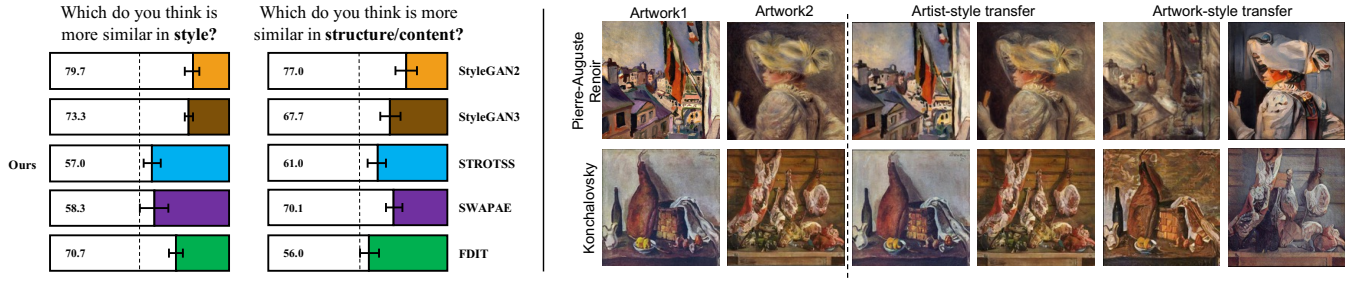

Figure 6: (Left) Human perceptual study on photo-realistic style transfer where we asked participants to choose which image better reflects the "style" or "content". Our model is rated best for capturing style and preserving content. (Right) Swapping the features of artist-style from the same artist but different artworks. The results show each distance of artist-style feature extracted from the same artist but different artworks through our method are close, and the generated images only have slight differences.

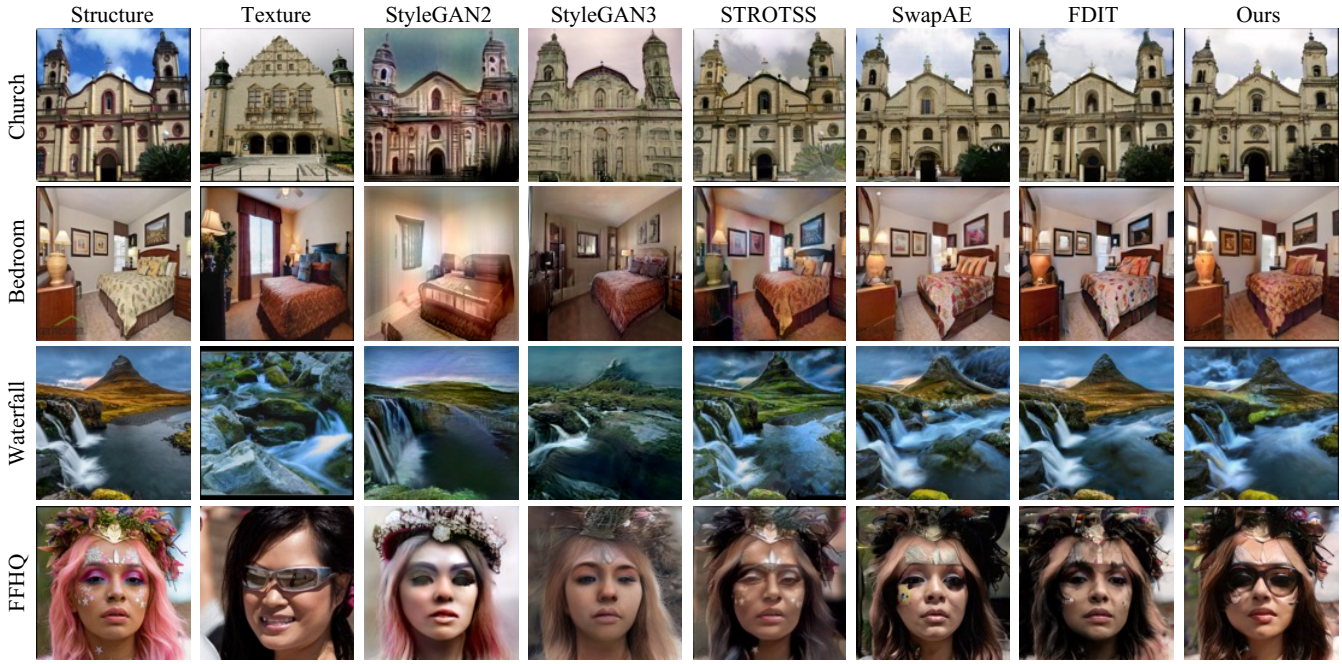

Figure 7: Photo-realistic style transfer comparison to evaluate the performance of preserving structural consistency. Results across four diverse benchmark datasets. Our approach generates realistic results and constrains more structural features.

Method	Train time (↓) (sec/epoch)	FID (↓) Church	FFHQ
w/o edge detection	16.85	52.34	59.83
w/ Canny detector	17.25	46.75	58.12
w/ FSE	26.24	48.93	58.96
w/o pooling	17.33	**42.66**	57.02
w/o reconstruction $\mathbf{z}_c$	17.28	45.89	57.85
Ours	17.34	42.71	**56.98**

Table 3: The ablation study on EEM. We compare two methods: Canny (1986) and FSE (Dollár and Zitnick 2014) with our method. The performance on w/o average pooling and w/o reconstruction $\mathbf{z}_c$ are in the 4th and 5th rows.

the average FID score from 48.74 to 42.83 across five large-scale benchmark datasets (Tab. 2).

LPIPS. LPIPS (Zhang et al. 2018) is the perceptual similarity of deep features extracted from two generate images with the same layout. We use LPIPS with a pre-trained AlexNet (Krizhevsky, Sutskever, and Hinton 2012) to measure the reconstruction quality between the original and generated images in Tab. 2. Our reconstructions better preserve the detailed outline than other baselines.

To summarize, our method preserves structural consistency of content images and reduces the introduction of structural features in style images. According to the user study, our method achieves both remarkable visual quality and satisfying style controllability.

Ablation Studies

Edge enhancing module. Here, we compare the FID and LPIPS values under different conditions in Tab. 3, including using different edge detection methods in EEM, with and without the average pooling layer, and taking only the structural component $\mathbf{z}_c$ as the input of generator G. The results show that our method strikes a balance by taking time cost and performance into account. Each component we used benefits for preserving structural consistency.

Double-style transferring module. To test DSTM further, we swap artist-style features of artworks between the same artist to show whether artist-styles are correctly extracted from the holistic perspective (Fig. 6).

Conclusion

In this paper, we propose DSTM to solve the challenge that artist-style extraction lacks generalization for unseen artists. DSTM swaps the two styles and generates realistic images given arbitrary style combinations. It extracts different artist-style and artwork-style from different artwork and preserves the diversity between different artworks of the same artist. We further propose an EEM to preserve structural consistency by deriving multi-scale and multi-level edge features from the holistic image.

Acknowledgments

This is paper is funded by the National Key R&D Program of China (2018AAA0100703) and the National Natural Science Foundation of China (No. 62006208 and No. 62107035).

References

Anokhin, I.; Solovev, P.; Korzhenkov, D.; Kharlamov, A.; Khakhulin, T.; Silvestrov, A.; Nikolenko, S.; Lempitsky, V.; and Sterkin, G. 2020. High-resolution daytime translation without domain labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7488–7497.

Arbelaez, P.; Maire, M.; Fowlkes, C.; and Malik, J. 2010. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5): 898–916.

Cai, M.; Zhang, H.; Huang, H.; Geng, Q.; Li, Y.; and Huang, G. 2021. Frequency domain image translation: More photo-realistic, better identity-preserving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13930–13940.

Canny, J. 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6): 679–698.

Chandran, P.; Zoss, G.; Gotardo, P.; Gross, M.; and Bradley, D. 2021. Adaptive convolutions for structure-aware style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7972–7981.

Chen, H.; Zhao, L.; Wang, Z.; Zhang, H.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2021. DualAST: Dual style-learning networks for artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 872–881.

Cheng, M.-M.; Liu, X.-C.; Wang, J.; Lu, S.-P.; Lai, Y.-K.; and Rosin, P. L. 2019. Structure-preserving neural style transfer. *IEEE Transactions on Image Processing*, 29: 909–920.

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*, 11326–11336.

Dollár, P.; and Zitnick, C. L. 2014. Fast edge detection using structured forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8): 1558–1570.

Dumoulin, V.; Shlens, J.; and Kudlur, M. 2016. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*.

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.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27: 2672–2680.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30: 6626–6637.

Hinton, G. E.; and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *Science*, 313(5786): 504–507.

Huang, S.; Jin, X.; Jiang, Q.; and Liu, L. 2022. Deep learning for image colorization: Current and future prospects. *Engineering Applications of Artificial Intelligence*, 114: 105006.

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*, 1501–1510.

Huang, Z.; Zhang, J.; and Liao, J. 2019. Style Mixer: Semantic-aware Multi-Style Transfer Network. In *Computer Graphics Forum*, volume 38, 469–480. Wiley Online Library.

Jing, Y.; Mao, Y.; Yang, Y.; Zhan, Y.; Song, M.; Wang, X.; and Tao, D. 2022. Learning Graph Neural Networks for Image Style Transfer. *arXiv preprint arXiv:2207.11681*.

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

Karayev, S.; Trentacoste, M.; Han, H.; Agarwala, A.; Darrell, T.; Hertzmann, A.; and Winnemoeller, H. 2013. Recognizing image style. *arXiv preprint arXiv:1311.3715*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34: 852–863.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119.

Kim, H.; Jhoo, H. Y.; Park, E.; and Yoo, S. 2019. Tag2pix: Line art colorization using text tag with secat and changing loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9056–9065.

Kolkin, N.; Salavon, J.; and Shakhnarovich, G. 2019. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10051–10060.

Kotovenko, D.; Sanakoyeu, A.; Lang, S.; and Ommer, B. 2019a. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4422–4431.

Kotovenko, D.; Sanakoyeu, A.; Ma, P.; Lang, S.; and Ommer, B. 2019b. A content transformation block for image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10032–10041.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25: 1097–1105.

Li, C.; and Wand, M. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, 702–716. Springer.

Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; and Yang, M.-H. 2017. Universal style transfer via feature transforms. *Advances in Neural Information Processing Systems*, 30: 386–396.

Liu, Z.-S.; Wang, L.-W.; Siu, W.-C.; and Kalogeiton, V. 2022. Name Your Style: An Arbitrary Artist-aware Image Style Transfer. *arXiv preprint arXiv:2202.13562*.

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

Park, T.; Zhu, J.-Y.; Wang, O.; Lu, J.; Shechtman, E.; Efros, A.; and Zhang, R. 2020. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33: 7198–7211.

Qiao, X.; Hancke, G. P.; and Lau, R. W. 2022. Learning Object Context for Novel-View Scene Layout Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16990–16999.

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.

Svoboda, J.; Anoosheh, A.; Osendorfer, C.; and Masci, J. 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*, 13816–13825.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.

Ulyanov, D.; Lebedev, V.; Vedaldi, A.; and Lempitsky, V. 2016. Texture networks: Feed-forward synthesis of textures and stylized images. *arXiv preprint arXiv:1603.03417*.

Wang, H.; Li, Y.; Wang, Y.; Hu, H.; and Yang, M.-H. 2020. Collaborative distillation for ultra-resolution universal style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1860–1869.

Xie, S.; and Tu, Z. 2015. Holistically-Nested Edge Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1395–1403.

Yoo, J.; Uh, Y.; Chun, S.; Kang, B.; and Ha, J.-W. 2019. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9036–9045.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.