REST: A Resolution Preserving Network for Photorealistic Style Transfer via Semantic Distillation⋆,⋆⋆

Jing Huo[a], Zheng Gu[b,a], Jiulin Zhang[a], Xiangde Liu[a], Shiyin Jin[a], Pingzhuo Tian[c], Wenbin Li[a], Jing Wu[d], Yu-Kun Lai[d] and Yang Gao[a,e]

[a]State Key Laboratory for Novel Software Technology, Nanjing University, NanJing, 210023, Jiangsu, China
[b]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, 518060, Guangdong, China
[c]School of Computer Engineering and Science, Shanghai University, 200444 Shanghai, China
[d]School of Computer Science and Informatics, Cardiff University, Cardiff, CF103AT, Wales, United Kingdom
[e]School of NetWork Security and Information Technology, Yili Normal University, Yining, 835000, Xinjiang, China

ARTICLE INFO

Keywords:
Photorealistic style transfer
Resolution preserving
Knowledge distillation

ABSTRACT

Photorealistic style transfer aims to adapt the style of a reference image to a content image while preserving photorealism. Existing methods primarily rely on downsampling auto-encoders, which sacrifices spatial details due to the reduced feature map resolution and leads to artifacts. To address this problem, we propose a Resolution-preserving nEtwork for Style Transfer (REST) to maintain full spatial resolution during stylization. Our framework employs a resolution-preserving (RP) network (RPNet) that retains input resolution, effectively preserving fine details. Although RP features capture intricate spatial information, their reliance on high-resolution processing can limit semantic depth. To enhance feature representation, we introduce a semantic distillation module that transfers hierarchical patterns from VGG features into RPNet, ensuring balanced detail and semantic fidelity. Combined with advanced style transfer blocks, REST achieves photorealistic results with improved efficiency. Experiments show improved performance and faster processing over existing photorealistic style transfer approaches.

1. Introduction

Given two photos, with one serving as the content photo and the other as style photo, photorealistic style transfer aims to apply the style of a reference photo to the content image while preserving its original content. Unlike artistic style transfer(Gatys et al., 2016), photorealistic style transfer necessitates the full preservation of all content details. This process plays a crucial role in image production, color style creation, and editing, tasks that remain labor-intensive and time-consuming despite the assistance of professional tools.

Photorealistic style transfer builds upon artistic style transfer techniques but prioritizes content detail preservation. While Gatys *et al.* (Gatys et al., 2016) pioneered neural style transfer using VGG-19 (Simonyan and Zisserman, 2014), subsequent methods (Huang and Belongie, 2017; Li et al., 2017b) inherited its fundamental limitation: detail loss from pooling operations (Li et al., 2018; Luan et al., 2017). Subsequent advances try to address this problem through various approaches: photorealism regularization (Luan et al., 2017), unpooling with mask preservation (Huang and Belongie, 2017), wavelet operations (Yoo et al., 2019), spatial propagation networks (Li et al., 2019a), structure decomposition (Park et al., 2020), and bilateral learning (Xia et al., 2020). In the past two years, Diffusion-based approaches have further advanced photorealistic style transfer. ArtAdapter (Chen et al., 2024) employs multi-level style encoding to preserve semantics during arbitrary style transfer, while Puff-Net (Zheng et al., 2024) introduces a lightweight content–style fusion framework for real-time stylization with fewer artifacts. More recently, SA-LUT (Gong et al., 2025) and ConsisLoRA (Chen et al., 2025) focus on efficiency and structural fidelity by leveraging spatially adaptive look-up tables and consistency-enhanced diffusion adapters, respectively. These advances highlight a growing emphasis on style fidelity. Despite these innovations, existing methods still produce noticeable artifacts: oversmoothing (as depicted in the results of PhotoWCT in Figure 1 (c)), blurriness and semantics errors (such as the results of WCT^2 and the representation of the sun's color in Figure 1 (d)) or local distortions (as illustrated by the patch results of WCT^2 in Figure 1 (f)).

In this paper, we identify that encoder downsampling inherently causes spatial detail loss (Zhao and Snoek, 2020), thus producing distortions in photorealistic style transfer. To resolve this, we propose REST - a REsolution-preserving network for photorealistic Style Transfer. Unlike existing approaches (Li et al., 2018; Yoo et al., 2019), REST employs a fully resolution-preserving encoder-decoder architecture that eliminates all pooling and upsampling operations. By maintaining full spatial resolution throughout processing, our network preserves fine image details more effectively than reduced-resolution VGG-based features.

Learning such resolution-preserving networks presents unique challenges. While preserving spatial details, these networks tend to overfit to low-level features, compromising structural semantics. To address this, we introduce semantic distillation, motivated by VGG-19's proven ability to capture high-level semantics despite its detail loss (Simonyan and Zisserman, 2014).

Our approach bridges this gap through semantic distillation, which effectively transfers VGG-19's hierarchical

Fig. 1. Given reference style images (a) and content images (b), we illustrate the photorealistic style transfer results of PhotoWCT (Li et al., 2018) (c), WCT^2 (Yoo et al., 2019) (d) and our model (e). In (f), enlarged patches of the content images and stylized results are given to compare image details. As can be seen, PhotoWCT suffers from over-smoothing effects and WCT^2 presents unclear fine details. In contrast, the proposed model produces better style transfer results with details better preserved and styles better transferred. Best view in large size to see details.

feature extraction capabilities to the resolution-preserving network while overcoming their dimensional mismatch via adaptive pooling and linear projection. By integrating this distilled knowledge with standard style transfer modules like AdaIN (Huang and Belongie, 2017) and WCT (Li et al., 2017b), the framework simultaneously achieves robust semantic stylization and unprecedented detail retention, outperforming existing methods in photorealistic quality.

Our main contributions can be summarized as follows:

First, we propose an innovative, resolution preserving network structure for effective photorealistic style transfer. It can also be extended to various other image processing tasks that necessitate image detail preserving, such as, image denoising.

Second, we introduce a semantic feature distillation module to train the resolution-preserving network. Dense layer-wise correspondences are established between the VGG network and the resolution-preserving network to transfer the VGG's semantic feature extraction capability to our proposed resolution preserving network.

Lastly, through extensive qualitative experiments, quantitative evaluations, and user studies, we demonstrate the efficacy of our approach in preserving content details while achieving effective stylization, outperforming state-of-the-art photorealistic style transfer methods. REST exhibits remarkable simplicity and high efficiency, achieving a 60× speed up over WCT^2.

2. Related Work

Since the first work of Gatys *et al.* (Gatys et al., 2016) which proposes an iterative neural optimization based method for artistic style transfer, a large amount of style transfer methods have been proposed. The main focus of these methods is to address the problems that iterative optimization is slow (Johnson et al., 2016; Sanakoyeu et al., 2018) and how to learn a model for universal style transfer (Jing et al., 2018; Dumoulin et al., 2017; Huang and Belongie, 2017; Li et al., 2019b, 2017a, 2018; Ulyanov et al., 2016, 2017; Hu et al., 2020; Chen et al., 2020a; Li, 2021; Yao et al., 2024; Cai et al., 2025). However, since these methods are designed for general artistic style transfer, they often result in spatial distortions (Xia et al., 2020), which are not ideal for photorealistic style transfer.

Photorealistic style transfer. Photorealistic style transfer has evolved along two main lines: traditional color/tone matching approaches and modern neural methods. Early work focuses on matching color statistics in restricted color spaces, achieving plausible global appearance transfer but limited structural control (Bae et al., 2006; Pitie et al., 2005; Reinhard et al., 2001). With deep learning, research has centered on mitigating geometric and structural distortion while improving realism (Luan et al., 2017; Yoo et al., 2019; Li et al., 2019b; Zhang et al., 2024; Chigot et al., 2025; Gong et al., 2025; Wang et al., 2024). Deep Photo Style Transfer (DPST) augments the optimization framework of Gatys et al. by introducing a local affine photorealism constraint that preserves the structure of the contentthe content-the contentthe contentthe contentthe contentthe contentthe content during stylization (Luan et al., 2017; Gatys et al., 2016). PhotoWCT replaces standard pooling/upsampling with pooling masks and unpooling to better maintain spatial correspondence, though a dedicated post-processing step is still required to smooth artifacts (Li et al., 2018). Building on whitening–coloring transforms, WCT2 substitutes pooling with wavelet pooling/unpooling, further improving photorealism and dispensing with post-processing (Yoo et al., 2019). To reinforce spatial affinity, Linear Style Transfer (LST) appends a spatial propagation network after the main stylization module (Li et al., 2019b). Xia et al. adopt a grid prediction module with a fully connected layer to learn local affine transforms that explicitly satisfy the photorealism constraint (Xia et al., 2020). Beyond these, Huo et al. proposed DIST, a lightweight image and video style transfer framework based on knowledge distillation and an efficient learnable feature transformation module, achieving high-quality results with significantly reduced model size(Huo et al., 2024). SA-LUT learns a spatially adaptive 4D look-up table conditioned on content/layout to realize photorealistic, edge-aware color/style mapping

with strong efficiency (Gong et al., 2025). ConsisLoRA enhances LoRA-based diffusion adapters via content- and style-consistency regularization, improving faithfulness to input structure while maintaining target style characteristics and supporting region-aware control (Chen et al., 2025). Overall, these methods constrain the stylized output to local affine color transforms derived from the input, but most are built upon auto-encoder architectures with downsampling, which inevitably degrade fine spatial details. In contrast, our approach designs a fully convolutional, resolution-preserving network that avoids downsampling and thus better maintains the spatial structure of the content. Extensive experiments demonstrate that our model preserves image details more effectively than existing methods.

Knowledge distillation (KD). Knowledge distillation (Hinton et al., 2015; Ba and Caruana, 2014; Buciluǎ et al., 2006) is a promising model compression technique that can transfer the knowledge of large networks (called the teacher) to small networks (called the student). Previously, KD is mainly explored in high-level vision tasks, such as image classification. Huan *et al.* (Wang et al., 2020) proposed a collaborative distillation for encoder-decoder based neural style transfer to reduce the number of convolutional filters. To learn a lightweight video style transfer network via knowledge distillation, Chen *et al.* (Chen et al., 2020b) used knowledge distillation to transfer knowledge from a large teacher video style transfer model to a smaller and more efficient video style transfer model. Huo *et al.* (Huo et al., 2024) employs knowledge distillation to compress the style transfer backbone, replacing the pre-trained VGG-19 with a lightweight student network while preserving feature representation capabilities for efficient style transfer.

In this paper, to force the proposed resolution preserving network to learn more useful features for style transfer, we proposed a semantic distillation module to transfer the high-level feature extraction ability of VGG-19 network to the resolution preserving network. Different from homogeneous student and teacher-based work (Wang et al., 2020; Chen et al., 2020b), we propose a method to distill knowledge from a heterogeneous teacher (VGG-19) to a heterogeneous student (RPEncoder).

3. Proposed Method

The overall structure of the proposed network is illustrated in Figure 2. The proposed model comprises three primary components: 1) A resolution-preserving network consisting of a resolution-preserving encoder (RPEncoder) and a resolution-preserving decoder (RPDecoder), as shown in Figure 2 (a). This network maintains the resolution of features throughout its forward process, thereby preserving input details. 2) To facilitate the RPEncoder in learning more effective features for style transfer, the semantic distillation module is proposed, see Figure 2 (b). This module takes as inputs the features from corresponding layers of both RPEncoder and the VGG-19 encoder. We employ adaptive pooling and linear projection to align the features initially, followed by forcing the two features to be close. 3) Lastly, a stylization block is also needed. In the proposed network structure of Figure 2 (c), we directly employ AdaIN as the fundamental stylization structure, augmented with multi-level stylization to augment the degree of stylization.

3.1. Resolution Preserving Network

The state-of-the-art photorealistic style transfer models (Li et al., 2018; Yoo et al., 2019) are mostly based on the auto-encoder structure, in which the resolutions are reduced by pooling operations. Reduction of the spatial dimension of feature maps may inevitably lead to loss of fine content details, which cannot be easily recovered by simple image reconstruction losses. To address this problem, we propose a resolution preserving network for photorealistic stylization, which contains a resolution preserving encoder and a resolution preserving decoder.

Resolution preserving encoder (RPEncoder). The RPEncoder, denoted as **E**, consists of several RPConvBlocks (see Figure 3). An RPConvBlock contains a 3×3 convolutional layer, a batch normalization (BN) layer, a rectified linear unit (ReLU) activation and a 1×1 convolutional layer. To avoid the loss of border pixels after convolution, the replication padding is employed before the operation of the 3×3 convolutional layer. For the 3×3 convolutional layer, the stride is set to 1. With the above design, the resolution of feature maps after RPConvBlocks will be preserved. Notice that to make the feature extraction ability of RPEncoder more powerful, we have also used a 1×1 convolutional layer. BN is used to control the scale of the output. Details of the network structure, including the number and the size of the convolutional filters, are given in the supplementary material.

Resolution preserving decoder (RPDecoder). With the features extracted by RPEncoder, we build the resolution preserving decoder **D**, called RPDecoder. Accordingly, it receives the full-resolution features from **E**, and learns to decode the output image. It is almost symmetrical with the RPEncoder, except that, each block of RPDecoder only consists of a 3×3 convolutional layer and a ReLU layer. BN and 1×1 convolutional layer are not included in RPDecoder.

Compared to the VGG-based structure, the resolution-preserving network (RPNet) receives each input with the same spatial resolution as original image. As resolution is preserved, it is noticed that the RPNet appears to consume significantly more memory and inference time than VGG-19.

Instead of that, as demonstrated in Tab. 2, our method exhibits a remarkable reduction in computational cost compared to previous methods. This is attributed to our method employing a network with multiple layers and generating features with significantly fewer channels.

3.2. Semantic Distillation

Utilizing the previously designed resolution preserving network, we can employ the same training scheme as that of the traditional auto-encoder to train the RPEncoder

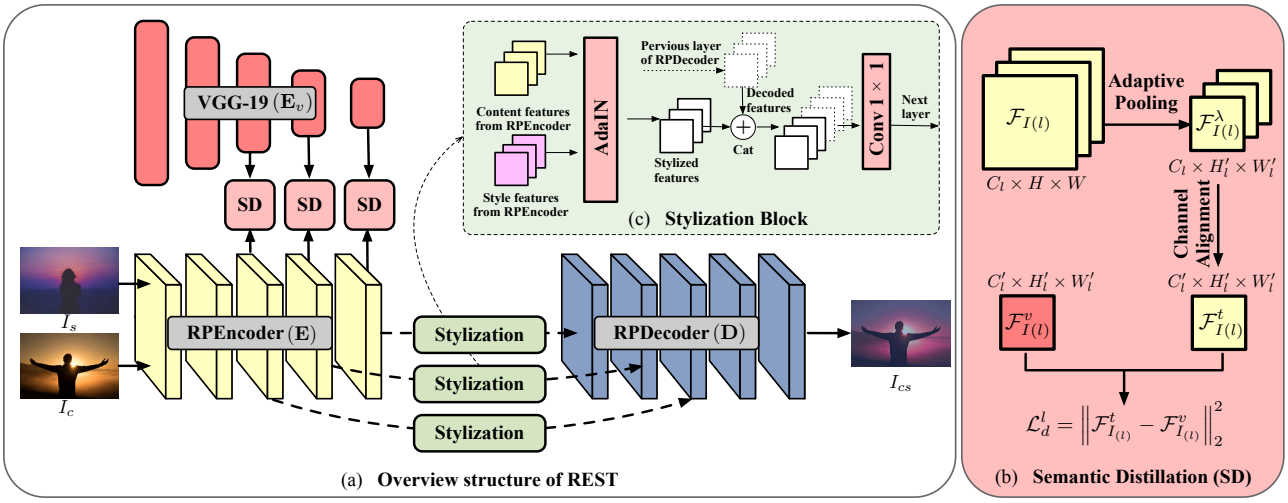


Fig. 2. Network Architecture. (a) An overview of our resolution preserving network for photorealistic style transfer (REST). The whole network contains a resolution preserving encoder (RPEncoder), a resolution preserving decoder (RPDecoder) and several stylization blocks. (b) In the training stage, semantic distillation (SD) is used to force the RPEncoder to learn high level semantic features which are similar to VGG features. (c) The stylization block is to produce stylized features via AdaIN-based feature transformation. Skip connections are used to enhance stylization degree.

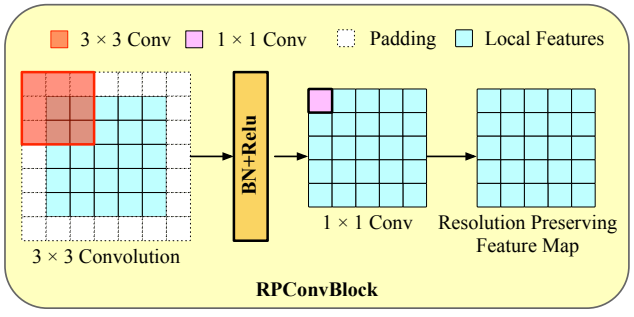


Fig. 3. The design of RPConvBlock. An RPConvBlock contains a 3×3 convolutional layer with stride 1 and replication padding, a BN layer, a ReLU layer and a 1×1 convolutional layer. It keeps the spatial resolution of the output feature map equal to its input's resolution.

and RPDecoder. However, as the features extracted by the RPEncoder are more informative than those obtained by the traditional encoder, the RPEncoder may inadvertently encode low-level features of the image, which we have observed to be less conducive to style transfer. Consequently, it is necessary to adapt the training schemes of the resolution-preserving network to facilitate the learning of high-level features that are more pertinent to colors, textures, and semantic meanings. Considering the successful use of VGG-19-based auto-encoders in style transfer, we introduce semantic distillation to transfer the feature extraction capabilities of VGG-19 to RPEncoder.

The semantic distillation module (SD) is shown in Figure 2 (b). Denote the VGG-19 encoder as $\mathbf{E}_v$. In our design, we set the layer number of RPEncoder the same as the VGG-19 encoder. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$ where W and H are the width and height of the input image, denote the output feature maps of the lth layer of RPEncoder and VGG-19 encoder as $\mathcal{F}_{I_{(l)}} \in \mathbb{R}^{C_l \times H \times W}$ and $\mathcal{F}^v_{I_{(l)}} \in \mathbb{R}^{C'_l \times H'_l \times W'_l}$. C_l and C'_l are the numbers of the lth-layer feature channels for RPEncoder and VGG encoder, and W'_l and H'_l are the feature map size of the VGG encoder. Because the RPEncoder preserves the full resolution and has fewer filters, its output feature size is different from the VGG output. To solve this problem and perform knowledge distillation, the following two steps are used to align the two features.

Spatial dimension alignment. For the first step, for an input feature pair $\mathcal{F}_{I_{(l)}}$ and $\mathcal{F}^v_{I_{(l)}}$, we align the spatial dimensions of the two features using adaptive average pooling. With the two feature maps' spatial sizes $H \times W$ and $H'_l \times W'_l$, we first calculate the adaptive pooling stride $\pi_{s(l)}$ and pooling window $\pi_{w(l)}$ accordingly and then use the average pooling operation.

$$\mathcal{F}^{\lambda}_{I_{(l)}} = \Lambda_{\pi_{s(l)}, \pi_{w(l)}} \left(\mathcal{F}_{I_{(l)}} \right) \tag{1}$$

where the size of the output $\mathcal{F}^{\lambda}_{I_{(l)}}$ is $C_l \times H'_l \times W'_l$. The $\Lambda_{\pi_{s(l)}, \pi_{w(l)}}(\cdot)$ denotes the average pooling operation with $\pi_{s(l)}, \pi_{w(l)}$ as parameters. It aggregates the spatial information of $\mathcal{F}_{I_{(l)}}$ to produce features. We observed that adaptive pooling operator produces smoother results than bicubic interpolation, in particular when the reference style image encodes diverse styles.

Channel dimension alignment. To further align the channel dimension of the two features, we consider that $\mathcal{F}^{\lambda}_{I_{(l)}}$ lies in a lower dimensional space of $\mathcal{F}^v_{I_{(l)}}$. Therefore, a linear projection operation is used to align the two feature dimensions.

$$\mathcal{F}^t_{I_{(l)}} = t(\mathcal{F}^{\lambda}_{I_{(l)}}) = Q \cdot \mathcal{F}^{\lambda}_{I_{(l)}} \tag{2}$$

where $Q \in \mathbb{R}^{C_l' \times C_l}$ is a transformation matrix. $\mathcal{F}^{\lambda}_{I_{(l)}}$ is flattened across spatial dimensions with size $C_l \times H_l' W_l'$. Then the new aligned features $\mathcal{F}^{t}_{I_{(l)}}$ have the same dimension as $\mathcal{F}^{v}_{I_{(l)}}$. During knowledge distillation, transformation matrix Q is learned together with the whole net using the semantic distillation loss.

Semantic distillation loss. With the two features aligned, the distillation loss at the lth layer can be formalized as mean squared loss:

$$\mathcal{L}_d^l = \|\mathcal{F}^{t}_{I_{(l)}} - \mathcal{F}^{v}_{I_{(l)}}\|_2^2 \tag{3}$$

We only apply the semantic distillation to the last three layers of the RPEncoder $\mathbf{E}$ and the VGG encoder $\mathbf{E}_v$, as shown in Figure 2 (b). This is due to the higher layers encode more high level (semantically related) representations than shallow layers. The overall distillation loss is summarized as follows:

$$\mathcal{L}_d = \sum_{l=L-2}^{L} \mathcal{L}_d^l \tag{4}$$

where L is the total number of layers, which equals 5 in our case. As the VGG encoder is trained on ImageNet (Deng et al., 2009) along with other classification tasks, it excels in extracting high-level semantic features. With the aforementioned design, the features of RPEncoder are now closely aligned with those of the VGG encoder, enabling us to transfer the semantic feature extraction capability from $\mathbf{E}_v$ to $\mathbf{E}$. In the future, we may also explore incorporating multi-task learning to compel the RPEncoder to learn semantic related features. It is worth nothing that, by enforcing the alignment of features $\mathcal{F}^{v'}_{I_{(l)}}$ to resemble $\mathcal{F}^{v}_{I_{(l)}}$, the original features $\mathcal{F}_{I_{(l)}}$ of RPEncoder contain both semantic information and fine image details preserved from the full resolution. This explains why our design is more suitable for photorealistic style transfer.

3.3. Stylization

The above resolution preserving network can be combined with various existing style transfer modules for photorealistic style transfer, such as AdaIN (Huang and Belongie, 2017) or WCT (Li et al., 2017b). However, since the WCT module is non-differentiable, gradients cannot be backpropagated when WCT is utilized. Therefore, in our study, we employ the AdaIN module to train our network in an end-to-end manner. In our experiments, we have also compared the performance with WCT using a fixed pretrained REST model. To incorporate AdaIN in our model, we have developed a noval multi-level progressive structure.

Progressive adaptive instance normalization. The AdaIN operation aligns the mean and the standard deviation of the content features $\mathcal{F}_{c_{(l)}}$ and the style features $\mathcal{F}_{s_{(l)}}$ using the following equation:

$$\begin{aligned}\mathcal{F}_{cs_{(l)}} &= \text{AdaIN}(\mathcal{F}_{c_{(l)}}, \mathcal{F}_{s_{(l)}}) \\ &= \sigma(\mathcal{F}_{s_{(l)}}) \left(\frac{\mathcal{F}_{c_{(l)}} - \mu(\mathcal{F}_{c_{(l)}})}{\sigma(\mathcal{F}_{c_{(l)}})} \right) + \mu(\mathcal{F}_{s_{(l)}})\end{aligned} \tag{5}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation operations calculated across spatial dimensions for each feature channel. However, in our experiments, we observed that applying AdaIN after the final block of the RPEncoder may result in less stylized ouycomess. As has also been found in several other works, multi-level style transfer has been shown to yield a more pronounced stylization effect. Therefore, we apply AdaIN at multiple layers of the encoder and tramsmit the stylized features to the decoder through skip connections. The stylized features are concatenated with the decoded features and then forwarded to the subsequent layer of decoder. This process is illustrated in Figure 2 (c). This progressive stylization process allows for the stylization degree in the output.

3.4. Overall Objective Function.

To train the above proposed model, we use the widely used content(Huang and Belongie, 2017) and style loss(Huang and Belongie, 2017; Li et al., 2017c), together with our proposed semantic distillation loss. Denote the output stylized image as I_{cs}. The content loss is defined at the 4th layer of the VGG-19 network.

$$\mathcal{L}_c = \|\mathcal{F}^{v}_{cs_{(4)}} - \mathcal{F}^{v}_{c_{(4)}}\|_2^2 \tag{6}$$

where $\mathcal{F}^{v}_{cs_{(4)}}$ and $\mathcal{F}^{v}_{c_{(4)}}$ denotes the extracted 4th layer features of I_{cs} and the content image I_c using $\mathbf{E}_v$.

$$\begin{aligned}\mathcal{L}_s = &\sum_{i=1}^{L} \left\| \mu\left(\mathcal{F}^{v}_{cs_{(i)}}\right) - \mu\left(\mathcal{F}^{v}_{s_{(i)}}\right) \right\|_2 + \\ &\sum_{i=1}^{L} \left\| \sigma\left(\mathcal{F}^{v}_{cs_{(i)}}\right) - \sigma\left(\mathcal{F}^{v}_{s_{(i)}}\right) \right\|_2\end{aligned} \tag{7}$$

where $\mathcal{F}^{v}_{s_{(i)}}$ denotes the extracted ith layer features of the style image I_s using $\mathbf{E}_v$. L is the total number of layers.

Given multiple input image pairs, with one image in the pair as the content and the other as style, the end-to-end model is optimized by the following multi-objective loss function:

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s + \lambda_d \mathcal{L}_d \tag{8}$$

where hyper-parameters λ_c, λ_s and λ_d control the contributions of each term. We empirically set $\lambda_c = 1$, $\lambda_s = 1$ and $\lambda_d = 2$ in all experiments.

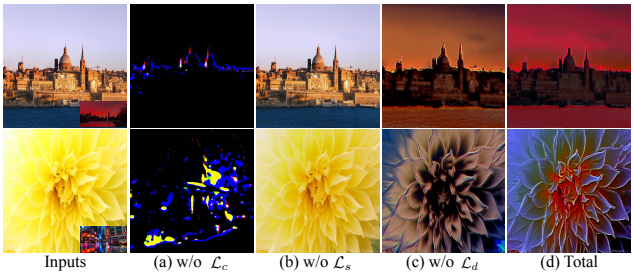
4. Experiments

In this section, we present ablation study results, followed by qualitative and quantitative comparison with state-of-the-art methods.

Implementation details. The pretrained VGG-based encoder is used with fixed weights. We trained our network using images from MSCOCO (Lin et al., 2014) and WikiArt (Duck, 2016) as content and style data respectively, by minimizing the loss defined in Eq. (8). The test set is obtained from Luan *et al.* (Luan et al., 2017) and Xia *et al.* (Xia et al., 2020). We have also expanded the test set with some high-quality photos for evaluation. The layer numbers

(a) Inputs (b) AEST (baseline) (c) REST (d) Patches

Fig. 4. Photorealistic style transfer results of resolution preserving network (REST) and auto-encoder (AEST).



Inputs (a) w/o $\mathcal{L}_c$ (b) w/o $\mathcal{L}_s$ (c) w/o $\mathcal{L}_d$ (d) Total

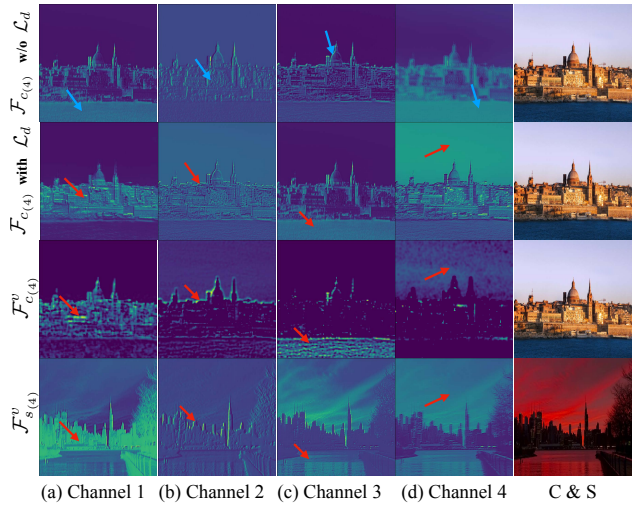
Fig. 5. Ablation study of the three main losses of REST. (a), (b) and (c) denote without $\mathcal{L}_c$, $\mathcal{L}_s$ and $\mathcal{L}_d$ respectively. (d) All losses are used.

L of the proposed encoder and decoder are both set as 5. We chose constant 16 channels for each layer. In each training batch, the data loader resizes one pair of content and style images with the size of 512×512 as input. The proposed model is implemented with PyTorch. The ADAM optimizer (Kingma and Ba, 2015) with a learning rate of 10^{-3} and batch size of 4 is used for training.

4.1. Ablation Study

Comparison with auto-encoder for photorealistic style transfer. We compare the stylization results of REST against the traditional auto-encoder based style transfer method (called AEST). We implemented AEST (our baseline) with analogous structure with REST, but with max pooling and upsampling after each block of **E** and **D**, respectively. As shown in Figure 4, since AEST reduces the spatial size of the feature maps, its results contain some checkerboard artifacts. In contrast, our REST model achieves pleasing results with content details preserved. This verifies the assumption that the resolution preserving mechanism is beneficial to photorealistic style transfer.

Influence of semantic distillation. The ablation results in Figures 5 (c) and (d) show that, $\mathcal{L}_d$ helps our model transfer correct color patterns. With $\mathcal{L}_d$ removed, our model tends to produce results less similar to the color pattern of the style images. To further explore this phenomenon, we visualize four channels of the feature maps of $\mathcal{F}_{c(4)}$ without and with $\mathcal{L}_d$, $\mathcal{F}^v_{c(4)}$ and $\mathcal{F}_{s(4)}$ from our RPEncoder and the VGG encoder respectively. Figure 6 shows the visualization results. As can be seen, it demonstrates that the semantic distillation contributes to helping RPEncoder capture similar semantic features as the VGG encoder. For example, in the 1st and 4th columns, $\mathcal{F}_{c(4)}$ with $\mathcal{L}_d$ and $\mathcal{F}^v_{s(4)}$ activate the same semantic region of architecture (first column) and sky (forth column), while $\mathcal{F}_{c(4)}$ without $\mathcal{L}_d$ activates the water region. Besides, according to the above observations, with the distillation loss, the AdaIN (Huang and Belongie, 2017) operation should be able to transfer color patterns among similar semantic regions by aligning channel-wise global statistic (mean and variance) of the content and style images. Results in Figure 5 (d), where the sky and architecture of the stylized result share similar style patterns with the corresponding regions of the style image, verify this.



(a) Channel 1 (b) Channel 2 (c) Channel 3 (d) Channel 4 C & S

Fig. 6. Visualization of feature maps. Green and black colors represent high and low responses of feature maps. The red arrows indicates that the networks activate image regions with same semantic meanings, while the blue arrows represent activated regions with misaligned semantic meanings.

(a) Input (b) L5 (c) L5 + L4 (c) L5 + L4 + L3

Fig. 7. Study of progressive stylization. (a) input content and style (bottom right corner) images. (b), (c) and (d) denote using stylization at the 5th layer, at the 4th-5th layers and at the 3rd-5th layers, respectively.

(a) Input (b) Use **AdaIN** (c) Use **WCT**

Fig. 8. Comparison of using AdaIN and WCT as style transfer modules.

Study of progressive stylization. We studied the influence of multi-level progressive stylization and results are

(a) Input (b) DPST (c) PhotoWCT (d) PhotoWCT(full) (e) LST

(f) LST (full) (g) WCT² (h) SWAE (i) InstantStyle (j) Ours (REST)

Fig. 9. Results of DPST (Luan et al., 2017), PhotoWCT (Li et al., 2018), LST (Li et al., 2019b), WCT^2 (Yoo et al., 2019), SWAE (Park et al., 2020), InstantStyle (Wang et al., 2024) and our method without mask guidance.

given in Figures 7 (b-c). As shown in the results, progressive stylization and feature fusion significantly improve the visual quality and stylization strength.

WCT *v.s.* AdaIN. For using WCT together with our network, we pretrained a resolution preserving network using reconstruction loss and distillation loss. As demonstrated in Figure 8, the results produced by AdaIN contain more fine details of the content image and higher sharpness than WCT. This is because RPEncoder is able to be trained in an end-to-end manner with AdaIN.

4.2. Visual Comparison

We evaluate the effectiveness of the proposed method by comparisons with state-of-the-art photorealistic style transfer methods, including DPST (Luan et al., 2017), PhotoWCT (Li et al., 2018), LST (Li et al., 2019b) and WCT^2 (Yoo et al., 2019). We obtained results using their official codes and default parameters settings. Two settings are adopted for comparison. In Figure 10, we utilized the segmentation masks provided by DPST during style transfer to compare mask-guided photorealistic style transfer results. In Figure 9, results without using masks are also provided.

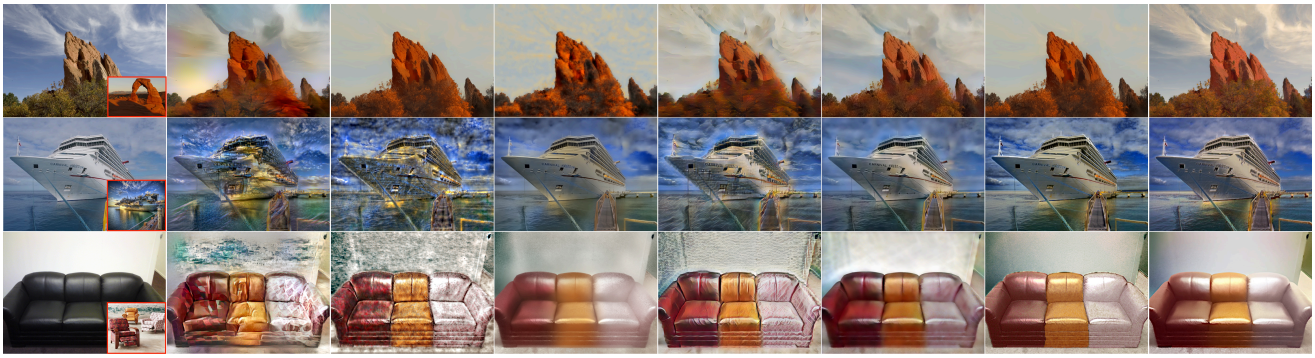


Fig. 10. Mask-guided photorealistic style transfer results of DPST(Luan et al., 2017), PhotoWCT(Li et al., 2018), LST(Li et al., 2019b) and WCT^2(Yoo et al., 2019) and our REST.

Table 1
Quantitative comparisons with state-of-the-art methods. "full" means post-processing technique of Li et al. (2018) is used. Higher SSIM score means the stylized image is more similar to the input content image in terms of fine details. Lower Gram loss denotes the stylized image presents more similar visual effects to the style image.

Method	DPST	PhotoWCT	PhotoWCT (full)	LST	LST (full)	WCT^2	AEST (baseline)	Ours (REST)
SSIM ↑	0.3295	0.4980	0.6097	0.3745	0.6590	0.6008	0.4334	**0.6218**
Gram Loss ↓	0.3888	0.8295	1.3005	0.5513	1.0757	0.9369	0.8602	**0.6872**

Table 2
Run time comparison of DPST, LST, PhotoWCT and WCT^2. We compare the computional time on content and style images with 512×512 resolution.

Methods	DPST	LST	PhotoWCT	WCT^2	Ours (REST)
Time(s)	113.85	0.96	8.78	1.84	**0.03**

Table 3
User study results. The percentages represent the ratio of how many times the method is selected as has fewest artifacts (**FS**), best stylization (**BS**) and best overall quality (**BQ**).

	DPST	PhotoWCT	LST	WCT^2	Ours (REST)
FS	11.84%	8.40%	18.30%	25.83%	**35.62**%
BS	**28.06**%	10.46%	15.31%	23.47%	22.70%
BQ	11.96%	9.99%	18.88%	27.55%	**31.61**%

Additionally, we present the results of PhotoWCT and LST without and with post-processing (PhotoWCT (full) and LST (full)). Note that results of our model (REST) do not involve any post-processing.

As shown in Figure 10, the results of DPST often exhibit distorted structures, which heavily hurt photorealism (e.g., Figure 10 (b) 2nd and 3rd rows). Although the results of PhotoWCT (full) have alleviated artifacts, they still suffer from over-smoothing and produce blurry outcomes due to the use of additional post-smoothing techniques to remove artifacts (e.g., Figures 10 (d) 2nd and 3rd rows). The results of LST also exhibit significant improvement after introducing post-processing. Similar over-smoothing effects are observed, as evident in Figure 10 (f), particularly noticeable in the sky in the 2nd row and the sofa in the 3rd row). WCT^2 employs wavelet pooling and unpooling in the style transfer network, enabling detail preservation. However, when used without semantic segmentation masks, WCT^2 may produce style transfer results with misaligned semantic regions as demonstrated in the 1st and 2nd rows of Figure 1 (g). We observed that WCT^2 failed to accurately transfer the sunset style. Figure 10 (h) presents the results of our method. In contrast, our proposed methods achieves effective photorealistic stylization and faithful detail preservation.

The above observations can also be found in Figure 9, we aim to verify that the proposed method is capable of transferring styles among similar semantic regions of the content and style even when masks are not used. In Figure 9 (h), SWAE (Park et al., 2020) can transfer textures well, but it heavily destroyed structure. JBL (Xia et al., 2020) produced similar results with ours. However, it performed unstable color transfer. For instance, in Figure 9 (i), the transferred color style in the first three images is weaker than in the last images. In contrast, our proposed method performs evidently better. This improvement is primarily attributed to our semantic distillation module, which aids in extracting high-level semantic-related features.

4.3. Quantitative Evaluation

Statistics. For further quantitative evaluation of our model, we use the structural similarity (SSIM) to assess the spatial structural similarity between the results and the original content. Additionally, we compute the Gram loss, as proposed by WCT^2 (Yoo et al., 2019), to evaluate the photorealistic stylization effects. Using a test set of 60 content-style image pairs (from Luan **et al.** (Luan et al., 2017)), we quantitatively evaluate the mask-guided photorealistic stylization of both our proposed method and

state-of-the-art approaches using SSIM and Gram loss metrics. Table 1 summarizes the results. It is evident that our proposed REST model achieved higher SSIM scores compared to our baseline (AEST), DPST, PhotoWCT, LST and WCT^2. It even surpasses the results of PhotoWCT with post-processing, indicating the superiority of our method in detail preservation. By comparing the results of PhotoWCT and LST with and without using post-processing, we found that the SSIM scores largely increased when post smoothing is applied. PhotoWCT and LST in fact heavily rely on using post-processing to achieve photorealism. Additionally, REST outperforms all compared methods in terms of Gram loss, except for DPST. We attribute this to the DPST directly optimizing the Gram loss. Nevertheless, our proposed method achieves fine detail preserving and faithful stylization despite its simple network structure.

Comparison of computational time. Table 2 a comparative analysis of computational efficiency between our proposed REST method and existing state-of-the-art approaches, revealing significant performance advantages in processing time. Conducted on a standardized hardware configuration featuring an Intel Xeon E5-2620 v4 processor and NVIDIA GeForce RTX 2080 Ti GPU under default experimental settings, the results demonstrate REST's remarkable computational speed, achieving a 60-fold improvement over WCT^2 and an extraordinary 3,790× acceleration compared to DPST. This dramatic performance enhancement stems from our carefully designed network architecture that combines structural simplicity with operational efficiency, maintaining high-quality stylization while drastically reducing computational overhead. The empirical evidence clearly establishes REST as the most efficient solution currently available for photorealistic style transfer tasks, setting a new benchmark for real-time performance in this domain.

User study. We conducted a user study to further validate whether our method achieves superior results compared to the baselines in terms of artifacts, stylization degree and overall quality. We recruited 18 users for the study. For every test, Each participant was presented with a content image, a style image, and several stylized results in random order. They were asked to pick three results, one with the fewest artifacts, one with best stylization degree and one with the best overall quality. To ensure consistency in evaluation, a short training session was conducted prior to the formal study. In this session, participants were provided with several example image pairs not included in the test set, together with explanations of the evaluation principles. Specifically, for artifacts (FS), participants were instructed to identify unnatural distortions such as blurriness, checkerboard patterns, color bleeding, or geometric inconsistencies, and select the result that minimized these issues. For stylization degree (BS), participants were asked to assess how faithfully the global appearance and local textures of the style image were transferred to the content image, with stronger and semantically aligned style adaptation considered preferable. For overall quality (BQ), participants were guided to balance both the absence of artifacts and the effectiveness of stylization, while also considering naturalness and visual plausibility of the result. After several practice comparisons illustrating these principles, participants confirmed their understanding before proceeding to the formal evaluation. In total, we collected 3240 responses (60 image pairs×3 questions×18 users) and calculated the percentage of each method selected as the best for the three evaluation aspects. Results are shown in Table 3. For all three evaluation aspects, REST outperforms the compared state-of-the-arts methods, except for stylization degree where DPST and WCT^2 are better. This maybe be attributed to DPST being an optimization-based method that conducts iterative optimization for every input image pair, resulting in better stylization degree. However, its results often contain distortion and artifacts (see Figure 9 (b) and Figure 10 (b)) . Our method demonstrates superior overall quality.

Discussion of failure cases. In some situations, we observed that the proposed method may exhibit strong edge effects at the border of objects when using masks as guidance. We attribute this issuse to the lack of accuracy in the masks. Additionally, the use of progressive stylization may exacerbate the impact of inaccurate borders. In the future, we intend to further explore and enhance our REST method for semantic region-aligned style transfer without relying on masks.

5. Conclusion

In this paper, we presented an end-to-end feed-forward resolution preserving neural network for photorealistic style transfer (REST). The key to our approach is learning meaningful resolution preserving features, which is done by semantic distillation to transfer the high level feature extraction ability to our network. The proposed network structure is further combined with a progressive stylization module for photorealistic style transfer. Extensive experiments in terms of visual, quantitative and inference time comparisons show that our model is significantly faster and better than the state-of-the-art methods. In future work, we aim to 1) achieve semantically aligned photorealistic style transfer in an unsupervised manner. 2) explore alternative methods for training the resolution-preserving network to extract high-level semantic features.

CRediT authorship contribution statement

Jing Huo: Conceptualization, Methodology, Writing– original draft. Zheng Gu: Conceptualization, Methodology, Writing– original draft. Jiulin Zhang: Software, Methodology, Writing– review editing. Xiangde Liu: Software, Conceptualization, Methodology, Writing– original draft. Shiyin Jin: Investigation, Validation, Writing– review editing. Pingzhuo Tian: Validation, Writing– review editing. Wenbin Li: Validation, Visualization, Writing– review editing. Jing Wu: Methodology, Supervision, Writing– review editing. Yu-Kun Lai: Methodology, Supervision, Writing–

review editing. Yang Gao: Supervision, Writing– review editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Data availability

Data will be made available on request.

Acknowledgements

This work is supported in part by National Natural Science Foundation of China (62276128, 62192783), Jiangsu Natural Science Foundation (BK20243051), Jiangsu Science and Technology Major Project (BG2024031), the Fundamental Research Funds for the Central Universities(14380128) and the Collaborative Innovation Center of Novel Software Technology and Industrialization, the Research Start-up Project of Shenzhen University (Grant No. RC20250382) and the Open Project of the State Key Laboratory for Novel Software Technology, Nanjing University (Grant No. KFKT2025B73), Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

Ba, J., Caruana, R., 2014. Do deep nets really need to be deep?, in: NeurIPS.

Bae, S., Paris, S., Durand, F., 2006. Two-scale tone management for photographic look. TOG 25, 637–645.

Buciluǎ, C., Caruana, R., Niculescu-Mizil, A., 2006. Model compression, in: SIGKDD, pp. 535–541.

Cai, B., Wang, H., Yao, M., Fu, X., 2025. Focus more on what? guiding multi-task training for end-to-end person search. IEEE Transactions on Circuits and Systems for Video Technology .

Chen, B., Zhao, B., Xie, H., Cai, Y., Li, Q., Mao, X., 2025. Consislora: Enhancing content and style consistency for lora-based style transfer. arXiv preprint arXiv:2503.10614 .

Chen, D.Y., Tennent, H., Hsu, C.W., 2024. Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8619–8628.

Chen, X., Yan, X., Liu, N., Qiu, T., Ni, B., 2020a. Anisotropic stroke control for multiple artists style transfer, in: ACM MM, pp. 3246–3255.

Chen, X., Zhang, Y., Wang, Y., Shu, H., Xu, C., Xu, C., 2020b. Optical flow distillation: Towards efficient and stable video style transfer, in: ECCV, Springer. pp. 614–630.

Chigot, E., Wilson, D.G., Ghrib, M., Oberlin, T., 2025. Style transfer with diffusion models for synthetic-to-real domain adaptation. arXiv preprint arXiv:2505.16360 .

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: CVPR, Ieee. pp. 248–255.

Duck, S.Y., 2016. Painter by numbers, wikiart. org. https://www.kaggle.com/c/painter-by-numbers.

Dumoulin, V., Shlens, J., Kudlur, M., 2017. A learned representation for artistic style, in: ICLR, OpenReview.net. URL: https://openreview.net/forum?id=BJO-BuT1g.

Gatys, L.A., Ecker, A.S., Bethge, M., 2016. Image style transfer using convolutional neural networks, in: CVPR, pp. 2414–2423.

Gong, Z., Wu, Z., Tao, Q., Li, Q., Loy, C.C., 2025. Sa-lut: spatial adaptive 4d look-up table for photorealistic style transfer. arXiv preprint arXiv:2506.13465 .

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 .

Hu, Z., Jia, J., Liu, B., Bu, Y., Fu, J., 2020. Aesthetic-aware image style transfer, in: ACM MM, pp. 3320–3329.

Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization, in: ICCV, pp. 1501–1510.

Huo, J., Kong, M., Li, W., Wu, J., Lai, Y.K., Gao, Y., 2024. Towards efficient image and video style transfer via distillation and learnable feature transformation. Computer Vision and Image Understanding 241, 103947.

Jing, Y., Liu, Y., Yang, Y., Feng, Z., Yu, Y., Tao, D., Song, M., 2018. Stroke controllable fast style transfer with adaptive receptive fields, in: ECCV, pp. 238–254.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution, in: ECCV, Springer. pp. 694–711.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, in: Bengio, Y., LeCun, Y. (Eds.), ICLR. URL: http://arxiv.org/abs/1412.6980.

Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z., 2019a. Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing 28, 492–505.

Li, R., 2021. Image style transfer with generative adversarial networks, in: ACM MM, pp. 2950–2954.

Li, X., Liu, S., Kautz, J., Yang, M.H., 2019b. Learning linear transformations for fast image and video style transfer, in: CVPR, pp. 3809–3817.

Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H., 2017a. Diversified texture synthesis with feed-forward networks, in: CVPR, pp. 3920–3928.

Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H., 2017b. Universal style transfer via feature transforms. NeurIPS 2017, 386–396.

Li, Y., Liu, M.Y., Li, X., Yang, M.H., Kautz, J., 2018. A closed-form solution to photorealistic image stylization, in: ECCV, pp. 453–468.

Li, Y., Wang, N., Liu, J., Hou, X., 2017c. Demystifying neural style transfer, in: Sierra, C. (Ed.), IJCAI, ijcai.org. pp. 2230–2236. URL: https://doi.org/10.24963/ijcai.2017/310, doi:10.24963/ijcai.2017/310.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: ECCV, Springer. pp. 740–755.

Luan, F., Paris, S., Shechtman, E., Bala, K., 2017. Deep photo style transfer, in: CVPR, pp. 4990–4998.

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

Pitie, F., Kokaram, A.C., Dahyot, R., 2005. N-dimensional probability density function transfer and its application to color transfer, in: ICCV, IEEE. pp. 1434–1439.

Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P., 2001. Color transfer between images. IEEE Computer graphics and applications 21, 34–41.

Sanakoyeu, A., Kotovenko, D., Lang, S., Ommer, B., 2018. A style-aware content loss for real-time hd style transfer, in: ECCV, pp. 698–714.

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

Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.S., 2016. Texture networks: Feed-forward synthesis of textures and stylized images., in: ICML, p. 4.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis, in: CVPR, pp. 6924–6932.

Wang, H., Li, Y., Wang, Y., Hu, H., Yang, M.H., 2020. Collaborative distillation for ultra-resolution universal style transfer, in: CVPR, pp. 1860–1869.

Wang, H., Spinelli, M., Wang, Q., Bai, X., Qin, Z., Chen, A., 2024. Instantstyle: Free lunch towards style-preserving in text-to-image generation. arXiv preprint arXiv:2404.02733 .

Xia, X., Zhang, M., Xue, T., Sun, Z., Fang, H., Kulis, B., Chen, J., 2020. Joint bilateral learning for real-time universal photorealistic style transfer, in: ECCV, Springer. pp. 327–342.

Yao, M., Wang, H., Chen, Y., Fu, X., 2024. Between/within view information completing for tensorial incomplete multi-view clustering. IEEE Transactions on Multimedia .

Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W., 2019. Photorealistic style transfer via wavelet transforms, in: ICCV, pp. 9036–9045.

Zhang, Z., Sun, J., Li, G., Zhao, L., Zhang, Q., Lan, Z., Yin, H., Xing, W., Lin, H., Zuo, Z., 2024. Rethink arbitrary style transfer with transformer and contrastive learning. Computer Vision and Image Understanding 241, 103951.

Zhao, J., Snoek, C.G., 2020. Liftpool: Bidirectional convnet pooling, in: ICLR.

Zheng, S., Gao, P., Zhou, P., Qin, J., 2024. Puff-net: Efficient style transfer with pure content and style feature fusion network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8059–8068.