Multi-Level Patch Transformer for Style Transfer with Single Reference Image

Yue He[1,2,†], Lan Chen[3,2,†], Yu-Jie Yuan[1,2], Shu-Yu Chen[1,2], and Lin Gao[1,2,*]

[1] Institute of Computing Technology, CAS
[2] University of Chinese Academy of Sciences
[3] Institute of Automation, CAS
†: Authors contributed equally * gaolin@ict.ac.cn

Abstract. Despite the recent success of image style transfer with Generative Adversarial Networks (GANs), this task remains challenging due to the requirements of large volumes of style image data. In this work, we present a deep model called *CycleTransformer* to optimize the mapping between a content image and a single style image by leveraging the strengths of transformer encoders and generative adversarial networks, where we advocate for patch-level operations. Our proposed network contains a Multi-level Patch Transformer encoder (MPT), which enables effective utilization of the style features of different scales. We combine the patch-based features with global feature maps to avoid overfitting to local style patterns, and feed them to a dynamic filtering decoder to adapt to different styles when generating the final result. Furthermore, we use a cycle-consistent training scheme to ensure the balance between content preservation and stylizing effects. Experiments and a user study confirm that our method substantially outperforms the state-of-the-art style transfer methods when both the style and content domain only contain one image each.

1 Introduction

Image Style transfer is a long-standing problem that seeks to convert an image to the style of a reference image. Despite the success of Generative Adversarial Networks (GANs) [13] in generating high-quality results, the requirement of an image database with the same style makes most of them cost prohibitive in real-world applications. To cope with a wide variety of unseen styles provided by common users that do not possess large datasets, some recent works [38, 33, 43] focus on modeling the internal statistics of patches contained in a content image and an arbitrary style image. They demonstrate the feasibility of exploiting only the information from the single content and style images by deep neural networks for the style transfer task.

Limited by the descriptive capability of the deep features learned by convolutional kernels with restricted local perceptual field and fixed structure, existing methods always generate undesirable stylized results when the reference style uses different patterns to depict the entire scene. Inspired by the powerful abilities of Transformer-based models [32, 12, 17, 7] on encoding rich relationships

existing in the input signals for various tasks, we consider that the self-attention mechanism is suitable to explore the internal relationships within the multi-scale image patches by capturing the long-distance dependencies among input elements. Another challenge for the style transfer task is the lack of paired training data. CycleGAN [49] has demonstrated to be effective in learning the style mappings between unpaired data by learning from large stylized image datasets. However, both our source and target domain only have one image each. To optimize the style mappings between two single unpaired images, we introduce a patch-level cycle-consistent learning scheme, which ensures high-quality stylization results that preserve the original semantic content.

In this work, we present *CycleTransformer*, a transformer-based neural method to deal with the style transfer from a single style image to a content image. CycleTransformer leverages patch-level self-attention and cross-level attention information for style mapping function optimization. More concretely, after extracting features using convolutional layers from randomly sampled nested patches of the input image, we not only learn the self-attention information within patches but also learn the attention across the patches of different scales in our novel Multi-level Patch Transformer encoder (MPT). MPT can exploit the possible relationships within and across the sampled patches to describe the style information carried by the input image. Since there is no prior restriction on the learning style, our network needs to cope with the large diversity in image styles, such as different sizes of color pieces and lengths of strokes, we integrate a dynamic filtering module [14] in our decoder to adaptively learn the filters for different styles when synthesizing the final results. Furthermore, To avoid getting stuck to certain local style modes, the learned embeddings by MPTs are combined with global features before being interpreted by decoding layers. When training our model, the cycle-consistent learning scheme is employed to optimize two mappings: mapping from the original image to the reference style and vice versa. Our network has the minimal requirement for common users in image stylization applications, and outperforms state-of-the-art methods on image-translation task when only one target domain image is available. Our main technical contributions are as follows:

- We propose a novel style transfer method with only one reference image.
- We propose a Multi-level Patch Transformer encoder to model the pixel-wise relationships within and across patches of multiple scales for effective patch-level style feature learning.
- A dynamic filtering module is applied for adapting to a broad range of image styles, that cooperates with our cycle-consistent learning framework to balance the content preservation and stylizing effects.

2 Related Work

Neural Style Transfer. Style transfer originated from non-photorealistic rendering [21] and has many applications such as natural image stylization [38, 39], augmented reality [1] and human face make-up [24], *etc.*Gatys *et al.*are the

pioneers who discovered that the features from a VGG-19 model lead to natural stylized results in the seminal works [8, 10], booming the development of optimization-based style transfer methods [22, 36, 25]. Instead of iterative optimization, feed-forward neural networks are proposed to accelerate the transfer process. Early works [18, 41, 23] train an independent network for each style, while the single network is further extended to multiple or arbitrary style transfer [2, 26, 6, 38, 28, 45] later on. Other approaches [29, 19, 31] are based on analogy or deformation, but require similar semantic structures between style and content images. According to the levels of stylization patterns, arbitrary style transfer can be further classified into two lines, *artistic stylization* and *photorealistic stylization*. The first line includes parameterized feature statistics by adaptive instance normalization (AdaIN) [16] and whitening and coloring transformation (WCT) [27]. To generate more sophisticated patterns, Sheng *et al.*[38] introduce a patch-based style decorator to reserve the detailed styles. The second line both gains stylization results and preserves photorealistic structural information via variants of WCT, *e.g.* coarse-to-fine recursive filtering [28] and wavelet corrected transfer network [46]. Alternatively, CycleGAN [49] uses a cycle consistency loss to constrain the GAN-based mappings from a source domain to a target domain and vice versa. However, CycleGAN needs large datasets to learn the bi-directional mappings. Park *et al.*[34] use patch-wise contrastive learning to do one-side image translation and can operate in a single image manner. Their method concentrates on the containing of structure information. Recently, the diffusion model [37] has achieved remarkable results in cross-modal image generation, and it has also been applied to image style transfer. For example, InST [47] uses textual inversion to obtain style-related embeddings from style images and then applies them to the conditional generation of stylized images. While our CycleTransformer aims to discover detailed style patterns with patch-wise attention in a cycle approach. Our CycleTransformer can generate both artistic stylization and realistic images in a single-image manner, and allow input images to have different style types and semantic structures.

Deep Vision Transformer In computer vision, researchers propose attention-based networks to capture long-range dependencies of pixels in images/videos which are beneficial to both classification [5] and regression [35, 17, 7], especially when coping with comparatively complex images. For image style transfer, some recent works utilize the self-attention mechanism to alleviate inductive bias caused by CNN kernel's priority for local interaction. For example, SANET [33] and AdaAttN [30] propose to learn correlations between the content and style feature maps by a learnable attention module. However, those feature transformation methods always fail to maintain content structures since they simply transfer features across all spatial locations for each channel. Yao *et al.*[44] add a transformer encoder into an autoencoder network to capture long-range region relations of the input image. But the patch-by-patch style swap and fusion in their method cause blurring. Deng *et al.*[4] introduce the transformer module into style transfer, taking image patches as words just like in NLP tasks, along with a progressive upsampling decoder to obtain clearer transfer results.

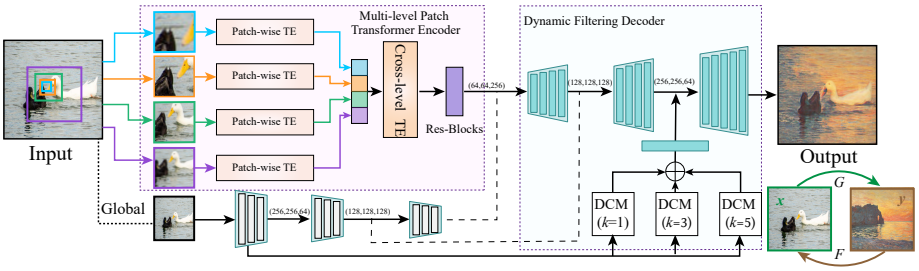


Fig. 1: The architecture of the generator of CycleTransformer. We use a multi-level Patch Transformer Encoder (MPT) to extract features from nested patches. Note that the sub-regions of the patch are randomly selected in the training stage. Thus it can robustly exploit different scales of contextual information at inference with randomly cropped patches. A dynamic filtering decoder interprets the patch-level and global features for result generation. We train two generators sharing the same structure with two discriminators in a cycle-consistent adversarial manner.

Our work uses a transformer-based approach to learn patch-wise style-related attention information for the image translation task.

3 Methodology

Given a content image x and a style image y, our goal is to transfer the style of y to x while maintaining the original semantic content. To achieve that for the two unpaired images, our network is designed to learn the bidirectional mappings between x and y to ensure that the generated image has both appropriate style features and preserves semantic content. Our network contains two mappings G and F sharing the same architecture, where G aims to transfer the style of y to x and F for vice versa. Fig. 1 shows the architecture of the generator of CycleTransformer. Instead of using one global image with a single cropped patch, we utilize the sub-regions of that patch to exploit different scales of contextual information as input. With extracted coarse- and fine-grained style features from *multi-level patch transformer encoders*, a *dynamic filter-based decoder* interprets the local and global features and produces the final translated image. During the inference stage, just randomly sampling one group of local nested patches with global features adequately generates good transfer results because the training stage exploits comprehensive contextual information.

3.1 Multi-level Patch Transformer Encoder

The distinguishable visual characteristics of different styles may exist within different spatial scales. For example, images with some painting styles may have

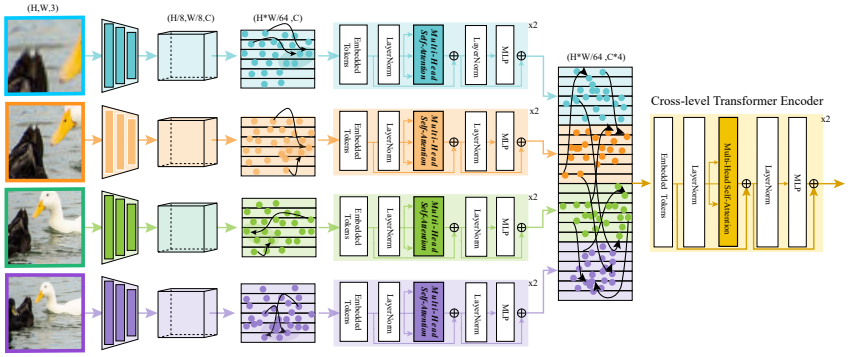


Fig. 2: The structure of Multi-level Patch Transformer (MPT) encoder. We use four patch-wise transformer encoders to learn the relationships within individual patches, and one cross-level transformer encoder to learn the relationships across different scales. The black arrows indicate the feature interactions based on the attention mechanism.

small pen strokes to express their content, while other styles could use much larger color blocks. To enhance the ability to capture visual patterns of different scales, we design a Multi-level Patch Transformer Encoder (MPT). The structure of MPT is shown in Fig. 2. The inputs of MPT are K nested patches of designated sizes, where we set K as 4 in all our experiments. The k-th patch p_k is firstly fed into convolutional layers to obtain its feature maps z^{p_k}, then the element-wise relationships within the patch feature maps are learned by a transformer encoder T, denoted by $t^{p_k} = T(z^{p_k})$. After learning the relationships inside each patch, we concatenate the transformed features and learn the attention information across all the patches through another transformer encoder. More specifically, the K cropped patches are resized to $256 \times 256 \times 3$ and fed to the downsampling sub-modules, each of which consists of one 3×3 convolution (padding=1), a normalization layer and a rectified linear unit (ReLU). We first convolve and downsample the feature maps twice and double the number of feature channels at each downsampling step. The size of the final feature map is $32 \times 32 \times 64$ (H/8, W/8, C). Directly feeding such a feature map to a transformer encoder would result in a huge memory cost. Instead, we unfold the feature map into a length-C sequence of $H \times W$-dimensional tokens to explore the channel-wise relationships in each patch encoder. The learned patch-level features t^{p_k} (k=1,2,3,4) are then concatenated in the channel dimension to generate a feature map of size $(H \times W/64, KC)$. Then we form the tokens representing features from different scales for the final transformer encoder to learn cross-level element relationships.

The structure of the transformer encoder [42] used in our network is shown in Fig. 2, which consists of $M = 2$ blocks containing a multi-head self-attention module and a feed-forward MLP layer. The positional encoding mechanism of Transformer strengthens its ability to exploit the relationships between the el-

ements by considering their relative positions. But the side effect of positional encoding is that if we only feed cropped patches to transformer encoders, the learned patch mapping would be too strong and directly convert the content image to the style image. To alleviate that problem, we incorporate the global features extracted from the entire image in all the decoding layers, since they provide the relationship between a cropped patch and the whole image so that the irrelevant relationships within patches can be weakened. Note that the global features are not utilized in MPT, as MPT is designed for investigating only patch-level features. If the global features are also included in MPT, MPT will tend to focus on the relationships between the patches and the fixed global structural features, leading to to a blurred result due to the lack of attention to style details.

3.2 Dynamic Filter-based Decoder

We introduce a dynamic filtering-based decoder to dynamically decide how the global features should be decoded when generating the final image. Previous works relying on convolutional layers with fixed kernels to decode features failed to infer different image styles adaptively [11, 9], as their filters cannot handle the large diversity of style-related features. Adapting to different kinds of styles is an essential requirement for the arbitrary style transfer task since we cannot assume any style prior before learning from the input images. Dynamic filtering module (DCM) [14, 48] simultaneously learns how to dynamically generate filtering kernels of different sizes for different input features. Then the deconvolution kernels applied on the feature maps can be customized based on the style of the input data. As shown in Fig. 1, we normally use 1, 3, and 5 as the sizes of the dynamically generated kernels to capture features of different scales. The three sub-modules to learn the dynamic filters are arranged in parallel. We feed the shallow global feature map to DCMs and concatenate the output with their input for the further decoding process.

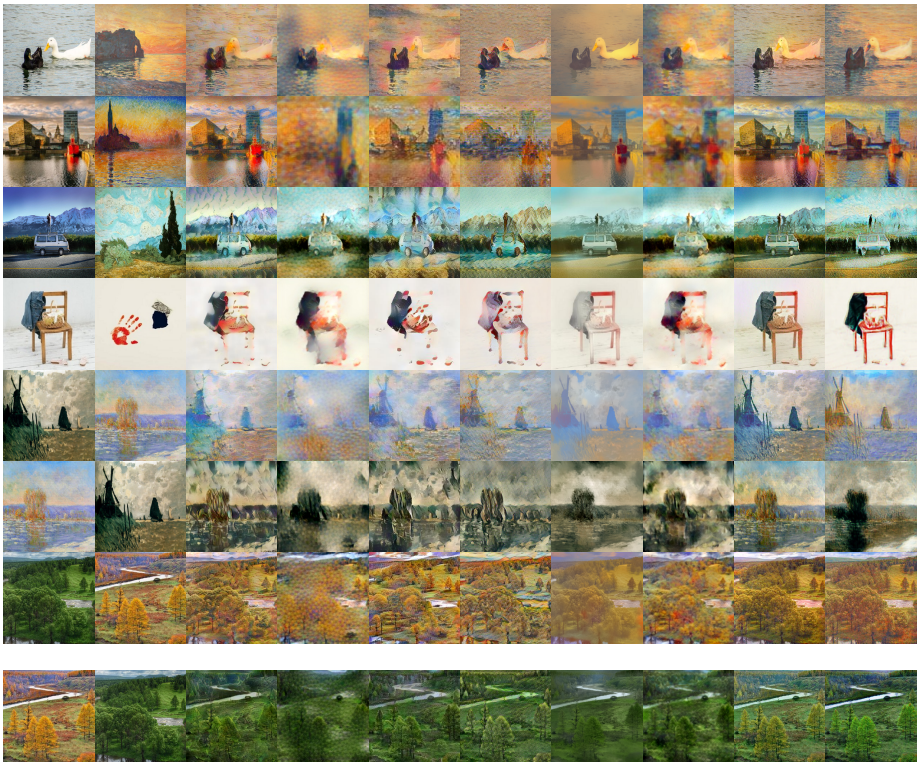
3.3 Loss Functions

Our method has two mapping functions to learn, $G : x \rightarrow y$ and $F : y \rightarrow x$. The discriminators and generators for G and F are trained under adversarial losses. The learning objective for G is:

$$\begin{aligned} \mathcal{L}_{GAN}(G, D_y, x, y) = E_y[logD_y(y)] \\ +E_{x,\{x(p_k)\}}[log(1 - D_y(G(x, \{x(p_k)\})))], \end{aligned} \tag{1}$$

where $x(p_k) \sim P_{data}\,[x(p)]$, representing the sampled patches from the data distribution $P_{data}\,[x(p)]$ of the patches with the same style of x. Similarly, for the mapping $F : y \rightarrow x$, we define the adversarial loss as $\mathcal{L}_{GAN}(F, D_x, x, y)$.

The adversarial losses are sufficient for generating plausible images in the target domain, but they cannot ensure the preservation of the original semantic content. Therefore, the two GANs need to be further updated using cycle



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 3: Results of the state-of-the-art algorithms and CycleTransformer on photo stylization. The first and second columns are content and style images. The remaining columns are stylized results by (c) AdaIN [16], (d) AAMS [44], (e) SANET [33], (f) StyleFormer [43], (g) photoWCT [28], (h) Avatar-Net [38], (i) DST [19] and (j) Ours.

consistency losses to encourage the synthesis of translations of the input image x. In the forward cycle, cycle consistency loss aims at translating the image $\widetilde{x}$ generated by G back to itself through F, which means using L_1 norm to measure the differences between $F(G(x, \{x(p_k)\}))$ and x. In the backward cycle, it calculates the differences between $G(F(y, \{y(q_k)\})$ and y, where $\{y(q_k)\}$ denotes the sampled patches from the data distribution $P_{data}[y(q)]$ with the style of y. The loss can be expressed as below:

$$\begin{aligned}\mathcal{L}_{cyc}(G, F, x, y) = E_{x,\{x(p_k)\}}[\|F(G(x, \{x(p_k)\})) - x\|_1] \\ + E_{y,\{y(q_k)\}}[\|G(F(y, \{y(q_k)\})) - y\|_1]\end{aligned} \tag{2}$$

To maintain the content of the content image, we add a reconstruction loss [49] to restrict the backward generation effect by L_1 loss. Since G is aimed at

enabling converting any content to the reference style of y, it should transform y back to itself, and so does the mapping F. The losses are defined as follows:

$$\begin{aligned}\mathcal{L}_{idt}(G, F, x, y) = E_{x,\{x(p_k)\}}[\|F(x) - x\|_1] \\ + E_{y,\{y(p_k)\}}[\|G(y) - y\|_1]\end{aligned} \tag{3}$$

In summary, the full objective function is:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{GAN}(G, D_y, x, y) + \mathcal{L}_{GAN}(F, D_x, y, x) \\ + \lambda_{idt}\mathcal{L}_{idt}(G, F, x, y) + \lambda_{cyc}\mathcal{L}_{cyc}(G, F, x, y),\end{aligned} \tag{4}$$

where the weights λ_{idt} and λ_{cyc} are set to 100.

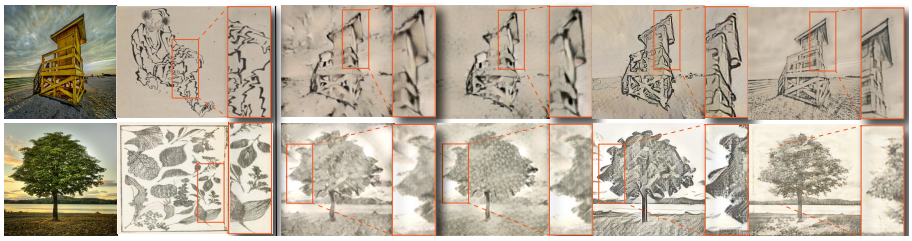
4 Experiments and Evaluations

4.1 Implementation Details

Our method only needs one content image and one style image to learn the bi-directional mappings, which facilitates our data collection. Most results presented in this paper are generated using images randomly selected from the CycleGAN dataset and downloaded from the internet. In our experiments, the four patch sizes are normally set as 4, 8, 16, and 32 respectively. All the cropped patches are first resized to 256*256 before being fed into the convolutional layers.

4.2 Qualitative Evaluation

We compare our method with state-of-the-art image stylization methods, including AdaIN [16], AAMS [44], SANET [33], StyleFormer [43], photoWCT [28], Avatar-net [38], DST [19], SpliceViT [40], StyTr^2 [4], QuantArt [15] and InST [47]. Our models are trained using only the two given images. We ran author-released implementations with their default settings for all the other methods. The first four rows in Fig. 3 are photo stylization results, where the content images are real photos and the style images do not contain similar semantic objects. The results of AdaIN [16] and Avatar-Net [38] cannot generate consistent patterns for similar regions, such as the sky regions in the fourth row and the building regions in the second row. PhotoWCT [28] fails to introduce the strokes or the color blocks to express the given styles in all the shown examples. AAMS [44] can integrate multiple stroke patterns and properly adopt the patterns in different regions of the output image. However, their image quality is not satisfactory due to the blurriness all over the picture. Since there is no obvious semantic correspondence between the two images, DST [19] transfers style like a color filter with a high retention of original appearances. In contrast, our method preserves the original content while producing appropriate style details. Cycle style transfer results are shown in the last four rows in Fig. 3 with artistic and photorealistic stylization. Since our network is powerful in learning positional relationships, and thus the style information extracted from style images can be



(a) Content (b) Style (c) Avatar-Net (d) AAMS (e) StyleFormer (f) Ours

Fig. 4: More comparisons on sketch styles. (a) and (b) are content and style images. We show the results of (c) Avatar-Net [38], (d) AAMS [44], (e) StyleFormer [43] and (f) Ours. It can be seen in the zoom-in windows that our method reveals the original stroke style in the best way.

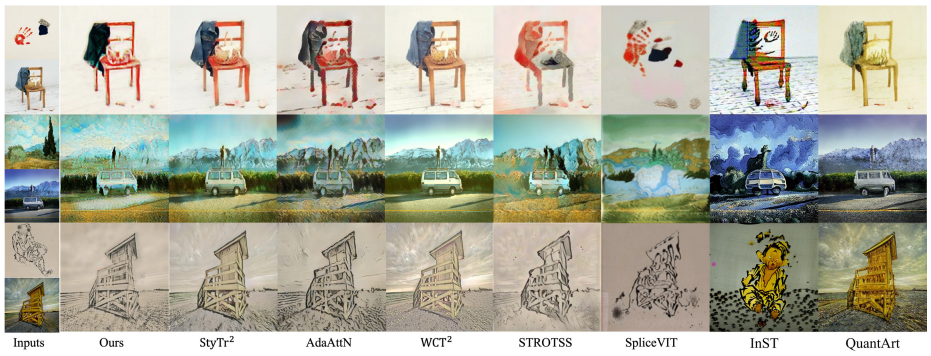


Fig. 5: More comparison results. Compared to StyTr^2 [4], AdaAttN [30], WCT^2 [45], STROTSS [20], SpliceViT [40], QuantArt [15] and InST [47], our method performs better in style transfer and content preservation.

properly distributed according to the content. The other methods either fail to learn the visual characteristics of the styles (AdaIN, photoWCT, and DST), or change the structure of the content (AAMS and Avatar-Net), leading to worse results than ours.

On the photo stylization for the challenging sketch styles (see Fig. 4), other methods are struggling to produce clean sketches and preserve the content structure properly at the same time. They leave original colors in the background except for photoWCT, which however produces blurred sketch lines. Our method is in a very advantageous position for this kind of style. In our results, only the grey-scale strokes from the style images are utilized to express the content, and the structural features are all well maintained. We also show more comparison results with StyTr^2 [4], AdaAttN [30], WCT^2 [45], STROTSS [20], SpliceViT [40], QuantArt [15] and InST [47] in Fig. 5, and our method performs better in style transfer and content preservation.

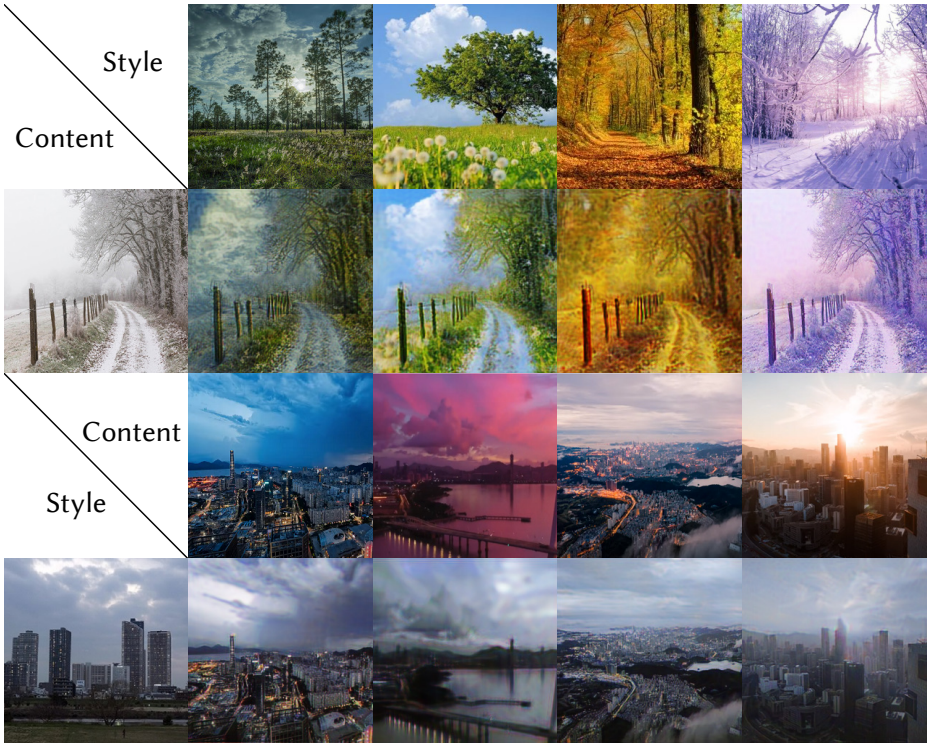


Fig. 6: Our photo-to-photo transfer results. We applied different styles to the same content image, and vice versa.

We demonstrate the ability of our methods in image translation between real-life photos. As shown in Fig. 6, our method can generate high-quality translated photos when the two images do not share semantically similar contents. By combining the global features with the features from dynamic filters as mentioned in Sec. 4.1, the details of the content images are well preserved. Fig. 6 shows the translated results when we use the images of four seasons as style image. By inspecting the four results of each example, we can see the content is consistent across the four seasons with properly altered appearances.

4.3 Ablation Study

We study the impact of different ingredients in our method and evaluate the structure of MPT. More experiments on the choice of hyper-parameters are shown in the supplementary materials.

CycleTransformer Architecture. Fig. 7 illustrates the generated results without MPT and/or Dynamic Filter Decoder. We take the CycleGAN trained by two single images (c) as our baseline. As CycleGAN is originally designed for

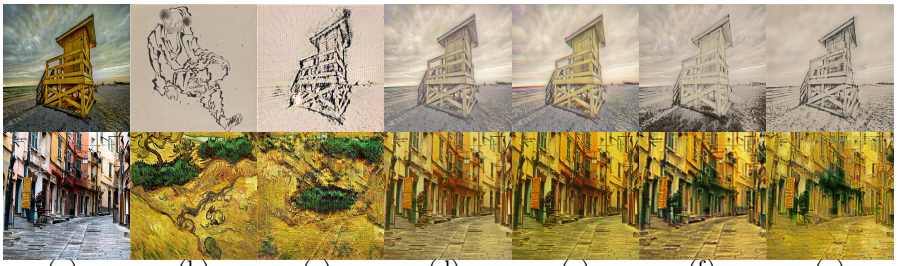


Fig. 7: Ablation study of the architecture of CycleTransformer. Given the content image (a) and the style image (b), we demonstrate the results where (c) the original CycleGAN is trained in a one-shot manner, (d) a random patch is added to extract convolutional features, (e) multiple patches are used to extract convolutional features, (f) multiple patches with dynamic filtering and our full model (g) with both MPT and dynamic filtering.

transferring a style learned from a large number of unpaired images, their one-shot results suffer from noises and blurriness. On the contrary, CycleTransformer achieves a much higher visual quality. To validate the benefits gained from the attention-based patch features extracted by MPT, we remove the MPT module and just feed one patch or nested patches to CNN layers to extract features and then combine them with global features to decode. We show the corresponding results in (d) and (e), where the generated blurred images both look like processed by a local color filter, which demonstrates the importance of the patch-wise self-attention and cross-level attention information learned by MPT when describing the intrinsic style-related features. We also validate the effectiveness of the Dynamic Filter Decoder, which adaptively interprets the shallow features to generate the final results. As shown in (f), the results generated by the model without dynamic filters have much weaker stylization effects. Compared with the results produced by the full model, they are less desirable due to the lack of artistic characteristics. The above study shows the irreplaceable roles of MPT and Dynamic Filter Decoder in CycleTransformer.

Multi-level Patch Transformer Encoder. We evaluate other alternative ways of using Transformer encoders and show the corresponding results in Fig. 8. If we simply use a Transformer encoder to process the global features, the network fails to learn delicate details and just generates weakly stylized results instead, as in Fig. 8(a). The reason is that it only learns the attention information globally and ignores the local pixel-wise relationships. We also evaluate the performance of only processing individual patches and directly feeding the concatenated multi-level patches to the cross-level Transformer encoder. As shown in the results of Fig. 8(b), if we learn only the self-attention information within patches, the network tends to apply a certain pattern universally, failing to adapt to the content. In the cross-level Transformer, the relationships between patches of different levels can be automatically revealed, which improves the stylization

effect as in Fig. 8(c). However, results from the full MPT are still superior to its results, because the relationships within the patches themselves are also essential for producing reasonable style details. Furthermore, MPT avoids the problem of duplicated patterns in some regions. As shown in the last column, using our MPT module, we can obtain rich style details with a reasonable distribution. Therefore, we can conclude that our MPT encoder module is able to effectively extract the relationships among pixels and patches of different scales for the arbitrary style transfer task.

Nested multilevel patches. Fig. 1 in the *supplementary* shows that nested multi-level patches produce better results than a single patch and the image quality improves with the number of nested patches. Compared with the ViT-based patch splitting schemes such as StyTr^2 (Fig. 5), our results show better brush textures. To further investigate the effects of "nested patches", we add an experiment where we replace the nested patches with the Gaussian pyramid of a randomly selected patch. Fig. 10 shows the results using patch pyramids, where the details are blurred with obvious artifacts which demonstrate the advantages of nested multi-level patches.

4.4 User Study

To evaluate the visual quality and the faithfulness of stylized images, we conducted a user study. We prepared 18 pairs of images, including 4 pairs in a cycle manner. We generated 22 groups of style transfer results using our method and seven state-of-the-art methods [16, 38, 44, 28, 19, 33, 43]. In total, 46 participants (including 29 males, 17 females, aged from 18 to 33) were recruited in this study and we got 46 (participants) $\times$ 22 (questions) = 1012 subjective evaluation results for each method.

The statistics of the user study results were plotted in Fig. 9. We performed one-way ANOVA tests on six methods with respect of "Style Consistency", "Content Consistency" and "Least Artifacts" corresponding to the three criteria above. We found significant effects of our method for all three criteria: style consistency ($F_{(5,126)} = 6.8, p < 0.0001$), content consistency ($F_{(5,126)} = 3.44, p < 0.05$) and least artifacts ($F_{(5,126)} = 2.36, p < 0.05$). Our method has obvious advantages in "Style Consistency", has a more consistent style with the reference image, and produces the least artifacts based on subjective evaluations. We also show a radar plot summarizing the user selections of the images with the best overall visual quality, where our method also got the most votes.

4.5 Quantitative Evaluations

We also sought image-translation tasks where we could get the ground truth translation results to evaluate our method quantitatively.

Visual Effect on both style and content We show quantitative comparisons with previous methods in Table 2 using the content/style perceptual losses (L_c and L_s) used in StyTr^2 [4] and FID on results generated using 300 style and content image pairs. WCT^2, STROTSS, and AdaIN gets the best L_c, L_s and

FID, respectively. Note that our method outperforms SpliceViT which has the same single image setting. In addition, we propose patchFID as a quantitative metric for style consistency (Table 2) since the artistic style features usually exist at the patch level. We randomly sampled 100 64×64 patches from the reference style image and regularly picked 16 64×64 patches from the stylized result, and took the mean of the FIDs between these patches as patchFID. We achieve the best patchFID. Note that there have not been any widely accepted metric for quantitative evaluation of style transfer, especially for artistic styles. We believe a user study that relies on human perceptual evaluation as in our paper is a meaningful and important measure.

Table 1: Quantitative evaluation.

Methods	Ours	StyTr^2	AdaAttN	WCT^2	STROTSS	SpliceViT	AdaIN	AAMS	SANET	StyleFormer	DST	photoWCT	Avatar-Net
L_c	2.31	1.00	2.55	**0.33**	2.19	3.01	1.76	2.50	2.46	2.41	0.69	1.16	2.12
$L_s \downarrow$	0.93	0.80	0.90	1.83	**0.34**	2.28	0.89	2.02	0.90	0.56	1.32	2.19	2.02
FID	2.77	1.73	5.81	1.26	5.42	8.58	**0.74**	4.12	2.33	1.74	4.88	7.64	7.83
patchFID $\downarrow$	**2.61**	3.21	2.78	4.05	3.33	7.62	3.74	3.72	4.57	3.31	3.49	5.20	6.17

Rationality of Color Distribution. We chose the colorization task since the paired grey-scale images and the colorized images are all available. We use the three datasets of different cartoon characters provided by [3] as our evaluation data. We randomly select 10 sets of unpaired color and grey-scale images as style and content images respectively from each of the three datasets. We train the models of CycleTransformer and the previous methods (*i.e.* AdaIN, Avatar-Net, AAMS, SANET, StyleFormer, photoWCT, DST, CycleGAN) in the same manner as mentioned in Sec. 4.2. Table 2 reports the performance of all the tested methods measured by the peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM). Our CycleTransformer significantly outperforms existing methods both in PSNR and SSIM. Fig. 11 shows some of the colorization results. It seems difficult for previous style transfer methods without cycle consistency (*i.e.* AdaIN, Avatar-Net, AAMS, photoWCT, DST) to learn complex patch-wise correspondences, leading to badly saturated colors. Compared to our method, CycleGAN is unable to achieve compelling results without being trained on large datasets.

4.6 Discussion CycleTransformer v.s. CycleGAN

CycleGAN is a successful approach for unpaired cycle image translation. But it still needs plenty of images for each domain in training. As shown in Fig. 7, their one-shot bi-directional mapping is unable to converge, thus the results suffer from noises and blurriness. The same artifacts can be seen in Fig. 11, a colorization task. Based on this work and the observed challenges, our work aims at using multi-level patch features for single reference image style transfer. For both photo stylization and colorization, CycleTransformer achieves a much higher visual quality on the contrary.

Table 2: Quantitative validation on the colorization task.

Method	character1		character2		character3	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Ours	22.327	**0.889**	**26.134**	**0.932**	**26.917**	0.920
AdaIN	16.724	0.809	17.315	0.863	17.865	0.850
AAMS	**22.626**	0.888	15.161	0.705	16.944	0.712
SANET	22.606	0.720	17.435	0.768	19.252	0.872
StyleFormer	13.243	0.433	12.338	0.705	15.324	0.657
DST	20.993	0.863	20.636	0.852	20.118	0.832
photoWCT	18.98	0.878	22.011	0.931	23.303	**0.928**
Avatar-Net	14.971	0.693	16.868	0.720	15.697	0.729
CycleGAN	12.484	0.472	14.762	0.472	12.754	0.556

5 Conclusion and Future Work

We focus on the challenge of learning to transfer style with only a single reference image. We introduce CycleTransformer, a deep model based on Transformers and the cycle-consistent learning scheme to model complex relationships within multi-level patches and across these patches. We integrate them with global features in a dynamic filter-based decoder to achieve a rich stylization effect and better content preservation. Our method uses randomly sampled patches to successfully model the distribution of the visual content with a certain style when only one style image is available. Experiments show the superiority of our method over the state-of-the-art methods on the single-image-based style transfer task. In the future, we will extend CycleTransformer to learn semantic-related patch-level features and utilize our feature learning scheme for other related tasks, such as sketch-based image synthesis.

Acknowledgments. This work was supported by the Beijing Municipal Natural Science Foundation for Distinguished Young Scholars (No. JQ21013), the National Natural Science Foundation of China (No. 62061136007) and the Youth Innovation Promotion Association CAS.

References

1. Castillo, C., De, S., Han, X., Singh, B., Yadav, A.K., Goldstein, T.: Son of zorn's lemma: Targeted style transfer using instance-aware semantic segmentation. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1348–1352. IEEE (2017)

2. Chen, D., Yuan, L., Liao, J., Yu, N., Hua, G.: Stylebank: An explicit representation for neural image style transfer. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2770–2779 (2017)
3. Chen, S.Y., Zhang, J., Gao, L., He, Y., hong Xia, S., Shi, M., Zhang, F.L.: Active colorization for cartoon line drawings. IEEE transactions on visualization and computer graphics **PP** (2020)
4. Deng, Y., Tang, F., Pan, X., Dong, W., Xu, C., et al.: Stytr2̂: Unbiased image style transfer with transformers. arXiv preprint arXiv:2105.14576 (2021)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv **abs/2010.11929** (2020)
6. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. ArXiv **abs/1610.07629** (2017)
7. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. arXiv preprint arXiv:2012.09841 (2020)
8. Gatys, L.A., Ecker, A.S., Bethge, M.: Texture synthesis using convolutional neural networks. In: NIPS (2015)
9. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)
10. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2414–2423 (2016)
11. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
12. Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: Video action transformer network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 244–253 (2019)
13. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
14. He, J., Deng, Z., Qiao, Y.: Dynamic multi-scale filters for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3562–3572 (2019)
15. Huang, S., An, J., Wei, D., Luo, J., Pfister, H.: Quantart: Quantizing image style transfer towards high visual fidelity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2023)
16. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017)
17. Jiang, Y., Chang, S., Wang, Z.: Transgan: Two transformers can make one strong gan. arXiv preprint arXiv:2102.07074 (2021)
18. Johnson, J., Alahi, A., Li, F.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision (ECCV). pp. 694–711 (2016)
19. Kim, S.S., Kolkin, N., Salavon, J., Shakhnarovich, G.: Deformable style transfer. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16. pp. 246–261. Springer (2020)
20. Kolkin, N., Salavon, J., Shakhnarovich, G.: Style transfer by relaxed optimal transport and self-similarity. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10051–10060 (2019)

21. Kyprianidis, J.E., Collomosse, J., Wang, T., Isenberg, T.: State of the âĂŸartâĂŹ: A taxonomy of artistic stylization techniques for images and video? (2012)
22. Li, C., Wand, M.: Combining markov random fields and convolutional neural networks for image synthesis. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2479–2486 (2016)
23. Li, C., Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks. ArXiv **abs/1604.04382** (2016)
24. Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W., Lin, L.: Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 645–653 (2018)
25. Li, Y., Wang, N., Liu, J., Hou, X.: Demystifying neural style transfer. arXiv preprint arXiv:1701.01036 (2017)
26. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Diversified texture synthesis with feed-forward networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 266–274 (2017)
27. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.: Universal style transfer via feature transforms. In: Advances in Neural Information Processing Systems (NIPS). pp. 386–396 (2017)
28. Li, Y., Liu, M.Y., Li, X., Yang, M.H., Kautz, J.: A closed-form solution to photorealistic image stylization. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 453–468 (2018)
29. Liao, J., Yao, Y., Yuan, L., Hua, G., Kang, S.B.: Visual attribute transfer through deep image analogy. ACM Trans. Graph. **36**(4), 120:1–120:15 (2017)
30. Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., Ding, E.: Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6649–6658 (2021)
31. Liu, X.C., Yang, Y.L., Hall, P.: Learning to warp for style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3702–3711 (2021)
32. Okamoto, T., Toda, T., Shiga, Y., Kawai, H.: Transformer-based text-to-speech with weighted forced attention. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6729–6733. IEEE (2020)
33. Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5880–5888 (2019)
34. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European Conference on Computer Vision. pp. 319–345. Springer (2020)
35. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N.M., Ku, A., Tran, D.: Image transformer. ArXiv **abs/1802.05751** (2018)
36. Risser, E., Wilmot, P., Barnes, C.: Stable and controllable neural texture synthesis and style transfer using histogram losses. arXiv preprint arXiv:1701.08893 (2017)
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
38. Sheng, L., Lin, Z., Shao, J., Wang, X.: Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8242–8250 (2018)

39. Texler, O., Futschik, D., Kučera, M., Jamriška, O., Sochorová, Š., Chai, M., Tulyakov, S., Sỳkora, D.: Interactive video stylization using few-shot patch-based training. ACM Transactions on Graphics (TOG) **39**(4), 73–1 (2020)
40. Tumanyan, N., Bar-Tal, O., Bagon, S., Dekel, T.: Splicing vit features for semantic appearance transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10748–10757 (2022)
41. Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.: Texture networks: Feed-forward synthesis of textures and stylized images. In: ICML (2016)
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
43. Wu, X., Hu, Z., Sheng, L., Xu, D.: Styleformer: Real-time arbitrary style transfer via parametric style composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14618–14627 (2021)
44. Yao, Y., Ren, J., Xie, X., Liu, W., Liu, Y., Wang, J.: Attention-aware multi-stroke style transfer. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1467–1475 (2019)
45. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic style transfer via wavelet transforms. In: International Conference on Computer Vision (ICCV). pp. 9035–9044 (2019)
46. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic style transfer via wavelet transforms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9036–9045 (2019)
47. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10146–10156 (June 2023)
48. Zhao, J., Chalmers, A., Rhee, T.: Adaptive light estimation using dynamic filtering for diverse lighting conditions. IEEE Transactions on Visualization and Computer Graphics **27**(11), 4097–4106 (2021)
49. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)

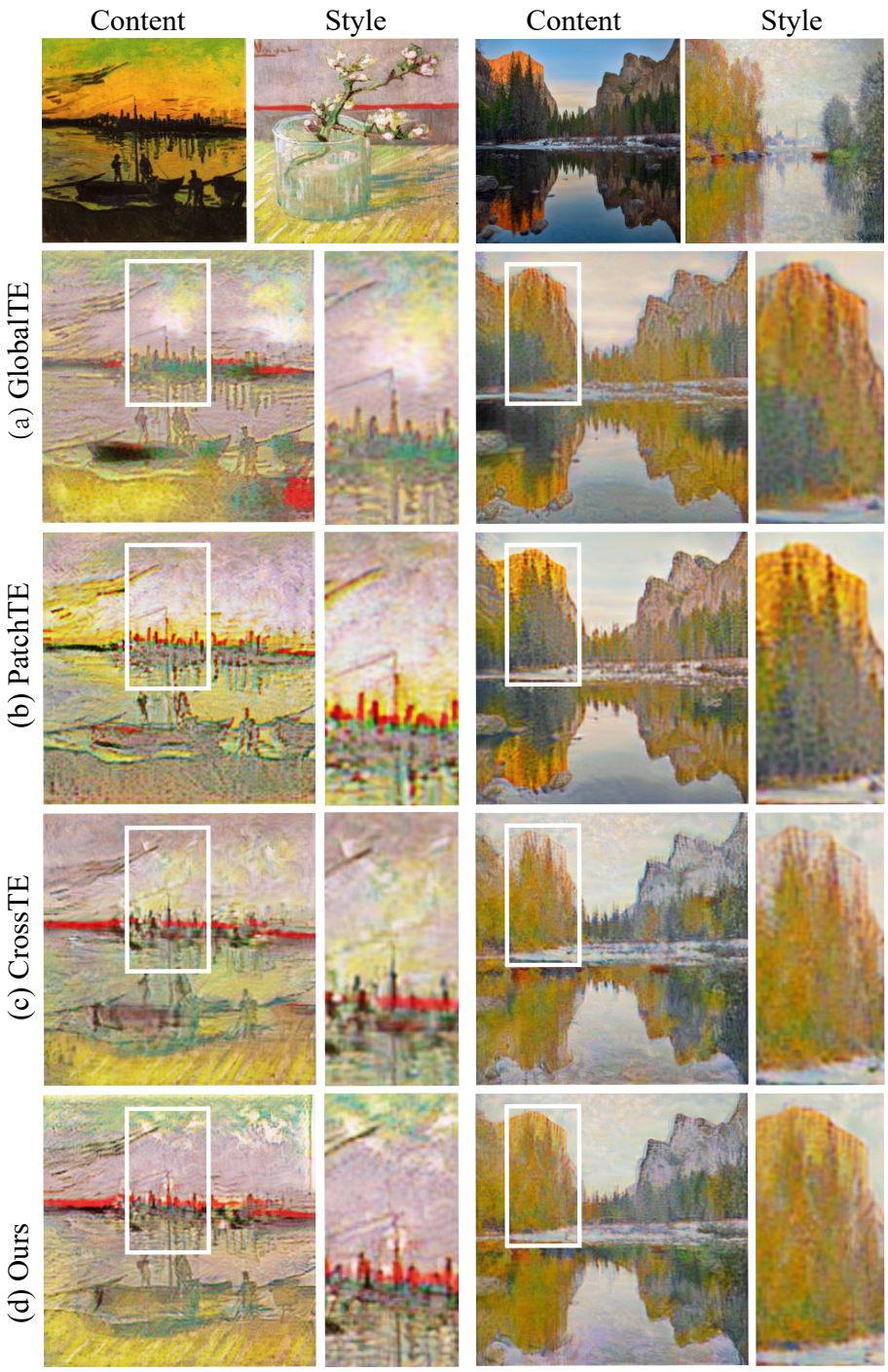


Fig. 8: Ablation study of MPT's structure. We compare the results of our method (d) with (a) "GlobalTE": only global features are fed to a transformer encoder (TE); (b)"PatchTE": patches are only processed by TEs individually; and (c) "CrossTE": patch features are directly processed by the cross-level TE.

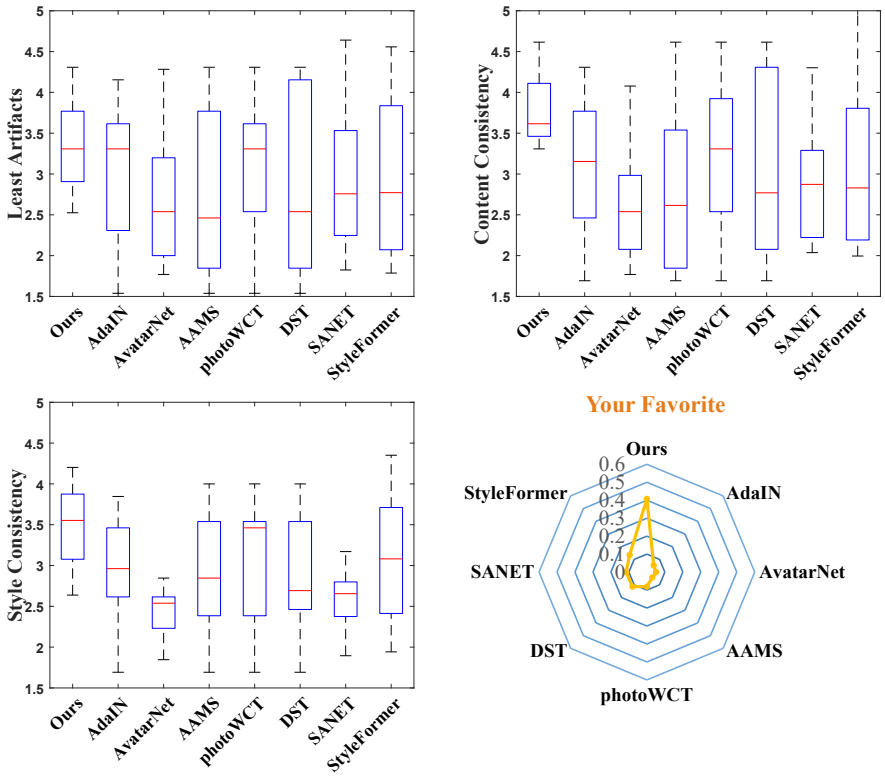


Fig. 9: Subjective scores on the similarity between the generated images of different methods and the corresponding content and style images, as well as the visual realism of the generated images. The additional radar plot summarizes the user selections of the images with the best overall visual quality.

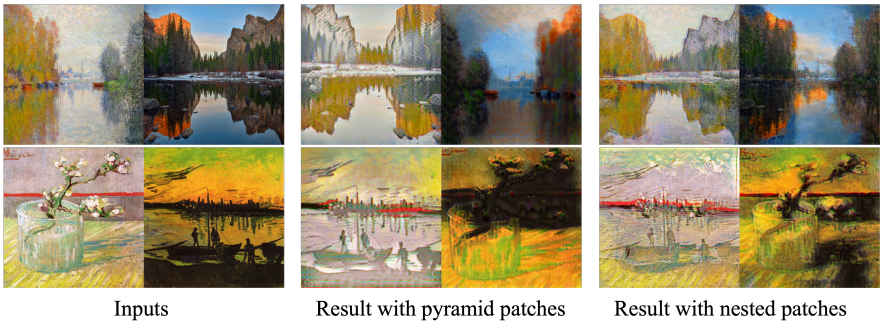


Fig. 10: Comparisons of patch pyramids and nested patches.

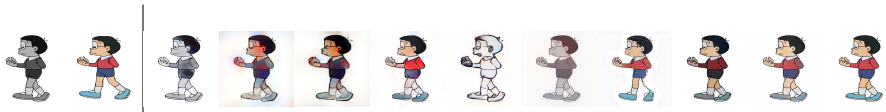


Fig. 11: Comparison of one-shot image colorization. (a) is the input grey image and (b) is the reference color image. The results of (c) AdaIN [16], (h) photoWCT [28], (g) StyleFormer[43] and (j) DST [19] generate results similar to global color filtering. The results of (d) Avatar-Net [38], (e) AAMS [44], (f) SANET[33] and (i) CycleGAN[49] suffer from blurry artifacts surrounding edges. Our method (k) is able to generate clean results and respect the boundaries in grey images well. (l) is the ground truth