Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Special Track on AI, the Arts and Creativity

DivSwapper: Towards Diversified Patch-based Arbitrary Style Transfer

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

Abstract

Gram-based and patch-based approaches are two important research lines of style transfer. Recent diversified Gram-based methods have been able to produce multiple and diverse stylized outputs for the same content and style images. However, as another widespread research interest, the diversity of patch-based methods remains challenging due to the stereotyped style swapping process based on nearest patch matching. To resolve this dilemma, in this paper, we dive into the crux of existing patch-based methods and propose a universal and efficient module, termed DivSwapper, for diversified patch-based arbitrary style transfer. The key insight is to use an essential intuition that neural patches with higher activation values could contribute more to diversity. Our DivSwapper is plug-and-play and can be easily integrated into existing patch-based and Gram-based methods to generate diverse results for arbitrary styles. We conduct theoretical analyses and extensive experiments to demonstrate the effectiveness of our method, and compared with state-of-the-art algorithms, it shows superiority in diversity, quality, and efficiency.

1 Introduction

Committed to automatically transforming the style of one image to another, style transfer has become a vibrant community that attracts widespread attention from both industry and academia. The seminal work of [Gatys *et al.*, 2016] first utilized the Convolutional Neural Networks (CNNs) to extract hierarchical features and transfer the style by iteratively matching the Gram matrices (i.e., feature correlations). Since then, valuable efforts have been made to improve the efficiency [Johnson *et al.*, 2016], quality [Lin *et al.*, 2021], and generality [Huang and Belongie, 2017], etc. However, as another important aspect of style transfer, *diversity* has received relatively less attention, and there are only a few works to solve this dilemma. [Li *et al.*, 2017b] and [Ulyanov *et al.*, 2017] introduced the diversity loss to train the feed-forward networks to generate diverse outputs in a learning-based mechanism. Alternatively, in a learning-free manner, [Wang *et al.*, 2020a] proposed to use Deep Feature Perturbation based on Whitening and Coloring Transform (WCT) to perturb the deep image feature maps while keeping their Gram matrices unchanged. These methods are all Gram-based; though considerable diversity can be achieved, unfortunately, they are not applicable to other types of approaches such as patch-based methods, since these methods are not underpinned by the Gram matrix assumption.

In this work, we are interested in the diversity of the patch-based stylization mechanism. As another widespread research interest of style transfer, the patch-based method is first formulated by [Li and Wand, 2016a; Li and Wand, 2016b]. They combined Markov Random Fields (MRFs) and CNNs to extract and match the local neural patches of the content and style images. Later, [Chen and Schmidt, 2016] proposed a Style-Swap operation and an inverse network for fast patch-based stylization. Since then, many successors were further designed for higher quality [Sheng *et al.*, 2018] and extended applications [Champandard, 2016], etc.

Let us start with a fundamental problem: *what limits the diversity of patch-based style transfer?* Whether using iterative optimization [Li and Wand, 2016a] or feed-forward networks [Chen and Schmidt, 2016], the core of patch-based methods is to substitute the patches of the content image with the best-matched patches of the style image (which we call "**style swapping**" [Chen and Schmidt, 2016] in this paper), where a *Normalized Cross-Correlation (NCC)* approach is mainly adopted to measure the similarities of two patches. However, as we all know, the NCC heavily depends on the consistency of local variations [Sheng *et al.*, 2018], and this stereotyped patch matching process restricts each content patch to be bound to its nearest style patch, thus limiting the diversity. Though it may be effective on semantic-level style transfer (e.g., portrait-to-portrait), for more general artistic styles (e.g., Fig. 1), there is little semantic correspondence between them and the contents. Even for human beings, it is hard to say which patches should match best. Therefore, we argue that for *artistic style transfer* [Gatys *et al.*, 2016], it would be more reasonable to relax the restricted style swapping process and allow some meaningful variations but maintain those inherent characteristics (e.g., the approximate semantic matching). It can give users more options to select the most satisfactory results according to different preferences.

*Corresponding authors.

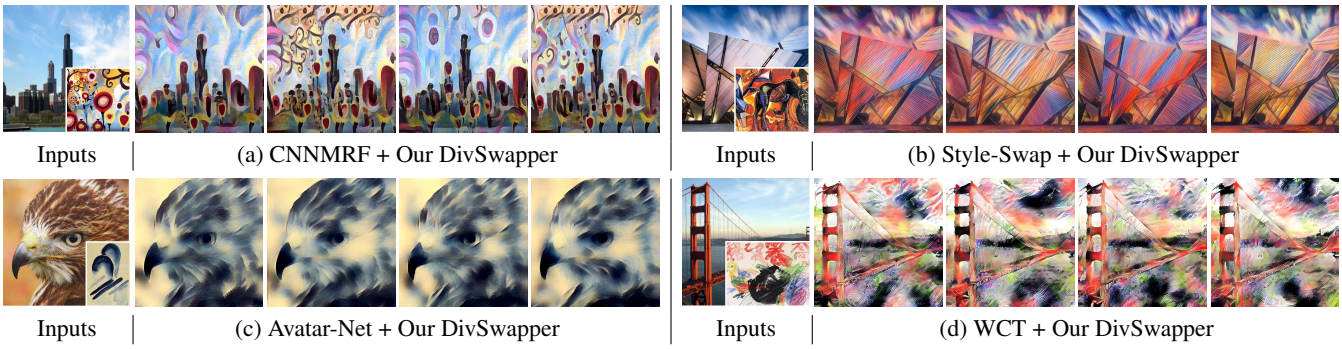

Figure 1: Given the same content and style images, our proposed DivSwapper can endow existing patch-based methods (e.g., (a) CNNMRF [Li and Wand, 2016a], (b) Style-Swap [Chen and Schmidt, 2016], and (c) Avatar-Net [Sheng *et al.*, 2018]) with the explicit ability to generate diverse stylized results. Moreover, it can also be intergated into Gram-based methods (e.g., (d) WCT [Li *et al.*, 2017c]) to achieve diversity. Our approach is plug-and-play and shows superiority in diversity, quality, and efficiency over the state of the art.

Moreover, for semantic-level style transfer, the diversified matching process can also help alleviate the undesirable artifacts caused by the restricted patch matching [Zhang *et al.*, 2019] (see later Sec. 4.3).

However, making such meaningful variations is a challenging task. First, neural patches are with high dimensions and hard to control. Maybe a small change would result in significant quality degradation, or a big change might not lead to a marked visual difference [Sheng *et al.*, 2018]. Therefore, the difficulty is finding the neural patches critical to visual variations and controlling them gracefully. Second, the visual effects and quality of the final results are also determined by the inherent correspondence between the content and style patches. Thus, how to manipulate this complicated correspondence to obtain diverse visual effects while maintaining the original quality is another problem to be solved.

Based on the above analyses, in this paper, we dive into the crux of patch-based style transfer and explore the universal way to diversify it. As shown in Fig. 2, an essential intuition we will use is that the visual effects of the output images are determined by the local neural patches of the intermediate activation feature maps; since the patches with higher activation values often contribute more to perceptually important (discriminative) information [Aberman *et al.*, 2018; Zhang *et al.*, 2018], they may also contribute more to visual variations as the human eyes are often more sensitive to the changes of these parts. In other words, if we could appropriately vary these higher-activated patches, then more significant diversity can be obtained. However, directly manipulating these patches is intractable since it is hard to distinguish which patches are with higher activation values and where they should be placed so as not to degrade the quality.

To remedy it, in this work, we theoretically derive that simply shifting the L2 norm of each style patch in the style swapping process can gracefully improve diversity and vary the patches with higher activation values in an implicit and holistic way. Based on this finding, we introduce a universal and efficient module, termed *DivSwapper*, for diversified patch-based arbitrary style transfer. Our DivSwapper is *plug-and-play* and *learning-free*, which can be easily integrated into existing patch-based methods to help them generate diverse outputs for arbitrary styles. Besides, despite building upon the patch-based mechanism, it can also be applied to Gram-based methods to achieve higher diversity (see examples in Fig. 1). Theoretical analyses and extensive experiments demonstrate that our DivSwapper can achieve significant diversity while maintaining the original quality and some inherent characteristics (e.g., the approximate semantic matching) of the baseline methods. Furthermore, compared with other state-of-the-art (SOTA) diversified algorithms, it shows notable superiority in *diversity*, *quality*, and *efficiency*.

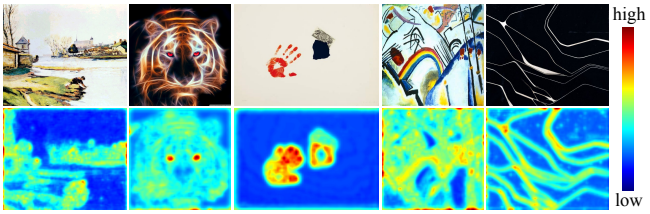

Figure 2: Our intuition: Patches with higher activation values often contribute more to perceptually important (discriminative) information such as semantics, salient colors, and edges; thereby, they could also contribute more to diversity. Top: Some style exemplars. Bottom: Heat maps of the activation feature maps (upsampled to the full image resolution) extracted from layer *Relu_4_1* of a pre-trained VGG19 [Simonyan and Zisserman, 2014].

Overall, the main contributions of our work are threefold:

- We explore the challenging problem of diversified patch-based style transfer and dive into its crux to achieve diversity. A universal and efficient module called *DivSwapper* is proposed to address the challenges and provide graceful control between diversity and quality.
- Our DivSwapper is *plug-and-play* and *learning-free*, which can be easily integrated into existing patch-based and Gram-based methods with little extra computation and time overhead.
- We analyze and demonstrate the effectiveness and superiority of our method against SOTA diversified algorithms in terms of diversity, quality, and efficiency.

Figure 3: Challenges of diversified patch-based style transfer. We adopt Avatar-Net [Sheng *et al.*, 2018] as the baseline method.

2 Related Work

The seminal work of [Gatys *et al.*, 2016] has ushered in an era of Neural Style Transfer (NST) [Jing *et al.*, 2019], where the CNNs are used to decouple and recombine the styles and contents of arbitrary images. After the recent rapid development, various methods have been proposed, among which the Gram-based and patch-based are the most representative.

Gram-based methods. The method proposed by [Gatys *et al.*, 2016] is Gram-based, using so-called Gram matrices of the feature maps extracted from CNNs to represent the styles of images, and could achieve visually stunning results. Since then, numerous Gram-based approaches were proposed to improve the performance in many aspects, including efficiency [Johnson *et al.*, 2016; Ulyanov *et al.*, 2016], quality [Li *et al.*, 2017a; Lu *et al.*, 2019; Wang *et al.*, 2020b; Wang *et al.*, 2021; Lin *et al.*, 2021; Chandran *et al.*, 2021; Cheng *et al.*, 2021; An *et al.*, 2021; Chen *et al.*, 2021b], and generality [Chen *et al.*, 2017; Huang and Belongie, 2017; Li *et al.*, 2017c; Li *et al.*, 2019; Jing *et al.*, 2020], etc.

Patch-based methods. Patch-based style transfer is another important research line. [Li and Wand, 2016a; Li and Wand, 2016b] first combined MRFs and CNNs for arbitrary style transfer. It extracts local neural patches to represent the styles of images and searches for the most similar patches from the style image to satisfy the local structure prior of the content image. Later, [Chen and Schmidt, 2016] proposed to swap the content activation patch with the best-matched style activation patch using a Style-Swap operation, and then used an inverse network for fast patch-based stylization. Based on them, many successors were further designed for better performance [Sheng *et al.*, 2018; Gu *et al.*, 2018; Park and Lee, 2019; Kolkin *et al.*, 2019; Yao *et al.*, 2019; Zhang *et al.*, 2019; Deng *et al.*, 2020; Liu *et al.*, 2021; Chen *et al.*, 2021a] and extended applications [Champandard, 2016; Liao *et al.*, 2017; Wang *et al.*, 2022].

Diversified methods. Our method is closely related to the existing diversified methods. [Li *et al.*, 2017b] and [Ulyanov *et al.*, 2017] introduced the diversity loss to train the feed-forward networks to generate diverse outputs by mutually comparing and maximizing the variations between the generated results in mini-batches. However, these methods are learning-based and have restricted generalization, limited diversity, and poor scalability [Wang *et al.*, 2020a]. To combat these limitations, [Wang *et al.*, 2020a] proposed a learning-free method called Deep Feature Perturbation to empower the WCT-based methods to generate diverse results. This method is universal for arbitrary styles, but unfortunately, it relies on WCT and does not apply to other types of methods.

Discussions. While there have been some efforts for the diversified style transfer, they are all Gram-based and are not applicable to other types of approaches such as patch-based methods. As another widespread research interest, the diversity of patch-based style transfer remains challenging. *Our work, as far as we know, takes the first step in this direction*. The proposed approach is learning-free and universal for arbitrary styles, and can be easily embedded into existing patch-based methods to empower them to generate diverse results. Moreover, it can also be applied to Gram-based methods to achieve higher diversity. Compared with the state of the art, our approach can achieve higher diversity, quality, and efficiency, which will be validated in later Sec. 4.

3 Proposed Approach

Before introducing our approach, let us first reiterate *why implementing diversity in patch-based methods is challenging?*

First, neural patches are with high dimensions and hard to control. On the one hand, maybe a small change would easily result in significant quality degradation, e.g., Fig. 3 (c). As can be observed in the red box areas, the result exhibits a severe quality problem that the portrait's eyes disappear, even if we only change 50 of the total 2500 neural patches. On the other hand, it is also possible that a big change may not lead to a marked visual difference, e.g., Fig. 3 (d). Although all the 2500 neural patches have been changed, the result is still very similar to the original one in Fig. 3 (b). Therefore, the difficulty is finding the neural patches *critical* to visual variations and controlling them *gracefully*.

Second, the inherent correspondence between the content and style patches ensures the visual correctness and rationality of the final results, as well as the semantic correspondence. If we simply ignore this local correspondence (e.g., randomly matching a style patch for each content patch), it will destroy the content prior and generate poor results, as shown in Fig. 3 (e). Therefore, one *key desideratum* of the diversified patch-based style transfer is to generate meaningful variations while maintaining the original quality and some inherent characteristics (e.g., the approximate semantic matching).

Aiming at the challenges above and based on the intuition introduced in Sec. 1, we propose a simple yet effective diversified style swapping module, termed *DivSwapper*, for diversified patch-based arbitrary style transfer. The proposed module is plug-and-play and learning-free, which can be easily integrated into existing patch-based and Gram-based methods to achieve diversity. As shown in Fig. 3 (f), the synthesized diverse results are all reasonable, with meaningful variations while maintaining the original quality and some inherent characteristics (e.g., the approximate semantic match-

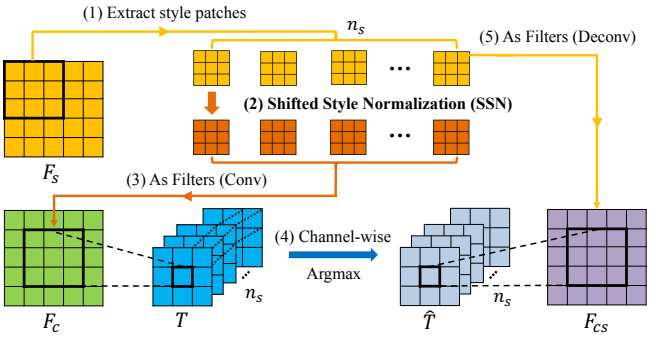

Figure 4: The workflow of our proposed DivSwapper.

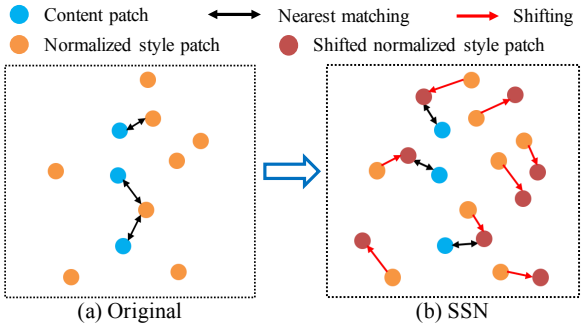

Figure 5: (a) The original style swapping process only produces one deterministic solution by matching the nearest style patch with each content patch. (b) Our SSN adds random deviations to shift the style patch normalization, thus achieving diversity.

ing). In the following section, we will first describe the workflow of our proposed DivSwapper, and then introduce the key finding and design in DivSwapper to achieve diversity, i.e., the *Shifted Style Normalization (SSN)*. In the light of this, we theoretically derive its effectiveness in generating diverse reasonable solutions and helping vary more significant neural patches with higher activation values.

3.1 Workflow of DivSwapper

Given a content image I_c and style image I_s pair, suppose $F_c = CNN(I_c)$ and $F_s = CNN(I_s)$ are content and style activation feature maps extracted from a certain layer (e.g., *Relu_4_1*) of a pre-trained CNN (e.g., VGG [Simonyan and Zisserman, 2014]). As shown in Fig. 4 (step (1-5)), our DivSwapper aims to search for the diverse yet plausible style patches in F_s for each content patch in F_c, and then substitute the latter with the former. The detailed workflow is:

(1) Extract the style patches from F_s, denoted as $\{\phi_j(F_s)\}_{j\in\{1,...,n_s\}}$, where n_s is the number of patches.

(2) Normalize each style patch by using a *Shifted Style Normalization (SSN)* approach. The shifted normalized style patches are denoted as $\{\hat{\phi}_j(F_s)\}$.

(3) Calculate the similarities between all pairs of the style and content patches by the *Normalized Cross-Correlation (NCC)* measure, i.e, $\mathcal{S}_{i,j} = \langle \phi_i(F_c), \hat{\phi}_j(F_s) \rangle$ (the norm of the content patch $\phi_i(F_c)$ is removed as it is constant with respect to the $\arg\max$ operation in the next step (4)). This process can be efficiently implemented by using a convolutional layer with the shifted normalized style patches $\{\hat{\phi}_j(F_s)\}$ as filters and content feature map F_c as input. The computed result T has n_s feature channels, and each spatial location is a vector of NCC between a content patch and all style patches.

(4) Find the nearest style patch for each content patch, i.e., $\phi_i(F_{cs}) = \arg\max_{j\in\{1,...,n_s\}} \mathcal{S}_{i,j}$. It can be achieved by first finding the channel-wise argmax for each spatial location of T, and then replacing it with a channel-wise one-hot encoding. The result is denoted as $\hat{T}$.

(5) Reconstruct the swapped feature F_{cs} by a deconvolutional layer with the *original style patches* $\{\phi_j(F_s)\}$ as filters and $\hat{T}$ as input.

Analysis: The most novel insight behind DivSwapper is that we use a *SSN* approach to inject diversity into the *NCC-based* style swapping process, which kills three birds with one stone: i) We can reshuffle *all* style patches by adding random norm shifts, which ensures the scope of diversity. ii) NCC is still used for nearest patch matching, and the final swapped feature F_{cs} is reconstructed by the *original style patches*, thereby the original quality and the inherent characteristics (e.g., the approximate semantic matching) can be well maintained. iii) SSN implicitly helps vary more significant style patches with higher activation values, thus achieving more meaningful diversity (see more analyses in Sec. 3.2). *Note that since the matching step (3) and the reconstruction step (5) actually can be implemented by two convolutional layers, our DivSwapper is very efficient.*

3.2 Shifted Style Normalization

The stereotyped style swapping process aims to search for the nearest style patch for each content patch, which only produces one deterministic solution, as illustrated in Fig. 5 (a). To obtain different solutions, an intuitive way is to match other plausible style patches instead of the nearest ones, which can be achieved by adjusting the distances between the content and style patches. However, as analyzed in Sec. 1, the key to obtaining more meaningful diversity is to gracefully control and vary those significant patches with higher activation values. Therefore, we propose the *Shifted Style Normalization (SSN)* to explicitly alter the distances between the content and style patches while implicitly restricting the swapping process to vary more significant style patches with higher activation values.

Simply yet non-trivially, as illustrated in Fig. 5 (b), our SSN adds a random *positive* deviation σ to shift the L2 norm of **each** style patch, like follows:

$$\{\hat{\phi}_j(F_s) = \frac{\phi_j(F_s)}{\| \phi_j(F_s) \| + \sigma}\}_{j\in\{1,...,n_s\}}. \quad (1)$$

Now, we theoretically derive the power of this *"magical"* random deviation σ to generate diverse solutions and help gracefully vary more significant style patches with higher activation values. For simplicity, we only take one content

and two style activation patches to illustrate, which are denoted as $\mathcal{P}^c$, $\mathcal{P}^s_1$, and $\mathcal{P}^s_2$, respectively. Note that the values in these vectors are *non-negative* because they are often extracted from the ReLU activation layers (e.g., *Relu_4_1*) of VGG model. Specifically, we first suppose that they satisfy the following original NCC matching relationship:

$$\frac{\langle \mathcal{P}^c, \mathcal{P}^s_1 \rangle}{\| \mathcal{P}^c \| \| \mathcal{P}^s_1 \|} = \cos\theta_1 > \frac{\langle \mathcal{P}^c, \mathcal{P}^s_2 \rangle}{\| \mathcal{P}^c \| \| \mathcal{P}^s_2 \|} = \cos\theta_2 > 0, \tag{2}$$

which means $\mathcal{P}^s_1$ matches $\mathcal{P}^c$ better than $\mathcal{P}^s_2$, where θ_1 is the angle between vector $\mathcal{P}^c$ and $\mathcal{P}^s_1$, θ_2 is the angle between vector $\mathcal{P}^c$ and $\mathcal{P}^s_2$. We want to change their matching relationship by randomly shifting the L2 norms of the style patches, i.e.,

$$\frac{\langle \mathcal{P}^c, \mathcal{P}^s_1 \rangle}{\| \mathcal{P}^c \| (\| \mathcal{P}^s_1 \| + \sigma_1)} < \frac{\langle \mathcal{P}^c, \mathcal{P}^s_2 \rangle}{\| \mathcal{P}^c \| (\| \mathcal{P}^s_2 \| + \sigma_2)}. \tag{3}$$

Thus, we can deduce:

$$\begin{aligned} &\langle \mathcal{P}^c, \mathcal{P}^s_2 \rangle \sigma_1 - \langle \mathcal{P}^c, \mathcal{P}^s_1 \rangle \sigma_2 > \\ &\langle \mathcal{P}^c, \mathcal{P}^s_1 \rangle \| \mathcal{P}^s_2 \| - \langle \mathcal{P}^c, \mathcal{P}^s_2 \rangle \| \mathcal{P}^s_1 \| . \end{aligned} \tag{4}$$

Since $\langle \mathcal{P}^c, \mathcal{P}^s_1 \rangle \| \mathcal{P}^s_2 \| - \langle \mathcal{P}^c, \mathcal{P}^s_2 \rangle \| \mathcal{P}^s_1 \| > 0$ (Eq. (2)), we can get the following solution:

$$\langle \mathcal{P}^c, \mathcal{P}^s_2 \rangle \sigma_1 - \langle \mathcal{P}^c, \mathcal{P}^s_1 \rangle \sigma_2 > 0 \Rightarrow \langle \mathcal{P}^c, \mathcal{P}^s_2 \rangle \sigma_1 > \langle \mathcal{P}^c, \mathcal{P}^s_1 \rangle \sigma_2.$$

As σ_1 and σ_2 are positive and i.i.d. (independent and identically distributed), it turns out that *holistically* our SSN *tends* to replace $\mathcal{P}^s_1$ with a suitable $\mathcal{P}^s_2$ which satisfies $\langle \mathcal{P}^c, \mathcal{P}^s_2 \rangle > \langle \mathcal{P}^c, \mathcal{P}^s_1 \rangle$. Since $\langle \mathcal{P}^c, \mathcal{P}^s_2 \rangle = \| \mathcal{P}^c \| \| \mathcal{P}^s_2 \| \cos\theta_2$, and $\langle \mathcal{P}^c, \mathcal{P}^s_1 \rangle = \| \mathcal{P}^c \| \| \mathcal{P}^s_1 \| \cos\theta_1$, we can deduce as follows:

$$\| \mathcal{P}^s_2 \| \cos\theta_2 > \| \mathcal{P}^s_1 \| \cos\theta_1 \Rightarrow \frac{\| \mathcal{P}^s_2 \|}{\| \mathcal{P}^s_1 \|} > \frac{\cos\theta_1}{\cos\theta_2}. \tag{5}$$

As $\cos\theta_1 > \cos\theta_2$ (Eq. (2)), we can obtain $\| \mathcal{P}^s_2 \| > \| \mathcal{P}^s_1 \|$, which means the varied $\mathcal{P}^s_2$ often has higher activation values than original $\mathcal{P}^s_1$. That is to say, our SSN could help vary more significant style patches with higher activation values in an implicit and holistic way. Besides, since it is still implicitly constrained by the original NCC (Eq. (2)) and the variations are gracefully controlled by the sampling range of σ, the overall quality and approximate semantic matching can be well preserved, as will be demonstrated in later Sec. 4.3.

4 Experimental Results

4.1 Implementation Details

Baselines. We integrate our DivSwapper into two types of patch-based methods based on (1) iteration optimization (CNNMRF [Li and Wand, 2016a]) and (2) feed-forward networks (Style-Swap [Chen and Schmidt, 2016] and Avatar-Net [Sheng *et al.*, 2018]). Besides, we also integrate it into a typical Gram-based method, i.e., WCT [Li *et al.*, 2017c]. We keep the default settings of these baselines and fine-tune the sampling range of our σ (sampled from a *uniform* distribution) to make our quality similar to the baselines, i.e., $(0, 10^3]$ for CNNMRF, $(0, 10^5]$ for Style-Swap, $(0, 5 \times 10^3]$ for Avatar-Net, and $(0, 5 \times 10^3]$ for WCT. We will discuss these settings in later Sec. 4.3. For more implementation details, please refer to the *supplementary material (SM)*.

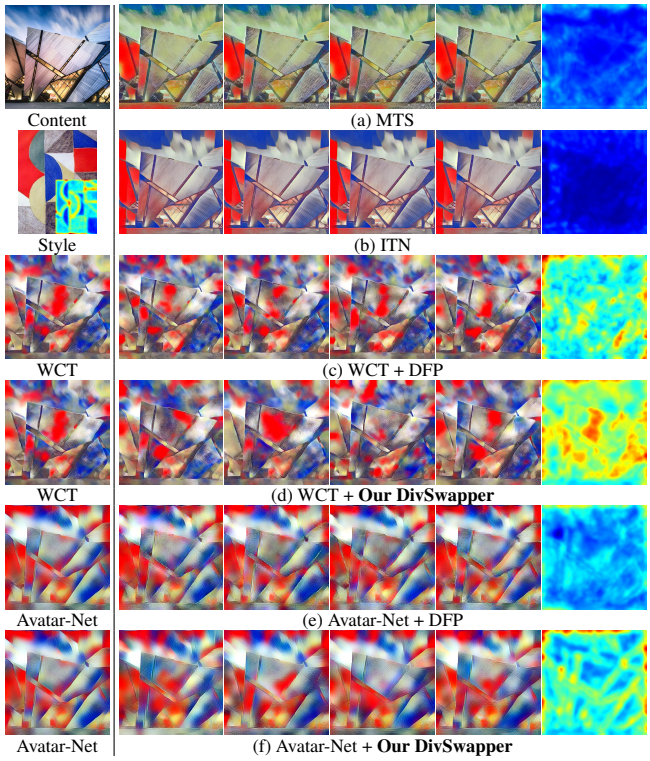

Figure 6: Qualitative comparisons. From top to bottom, the first column shows the input content and style images and the original outputs of baselines; the middle four columns show the diverse outputs of MTS, ITN, baselines + DFP, and baselines + our DivSwapper [*best viewed in color changes and zoomed-in*]. We also visualize their average activation feature differences (averaged on $C^2_{20} = 190$ pairs of diverse results) via heat maps in the last column.

Metrics. To evaluate the diversity, we collect 36 content-style pairs released by [Wang *et al.*, 2020a]. For each pair, we randomly produce 20 outputs, so there are a total of $36 \times C^2_{20} = 6840$ pairs of outputs generated by each method. Like [Wang *et al.*, 2020a], we adopt the average pixel distance D_{pixel} and LPIPS (*Learned Perceptual Image Patch Similarity*) distance D_{LPIPS} [Zhang *et al.*, 2018] to measure the diversity in pixel space and deep feature space, respectively.

4.2 Comparisons with Prior Arts

We compare our DivSwapper with three SOTA diversified methods, i.e., Multi-Texture-Synthesis (MTS) [Li *et al.*, 2017b], Improved-Texture-Networks (ITN) [Ulyanov *et al.*, 2017], and Deep-Feature-Perturbation (DFP) [Wang *et al.*, 2020a]. Since these methods are all Gram-based, we integrate our DivSwapper into the Gram-based baseline WCT [Li *et al.*, 2017c] and the Gram-and-patch-based baseline Avatar-Net [Sheng *et al.*, 2018] for a fair comparison.

Qualitative Comparison. As shown in Fig. 6 (a,b), MTS and ITN only achieve subtle diversity, which is hard to perceive. In rows (c,e), DFP can diversify WCT and Avatar-Net to generate diverse results, but the diversity is still limited, especially for Avatar-Net. On the same baselines, our DivSwapper achieves much more significant diversity, e.g., the

Figure 7: Our DivSwapper can diversify existing pure patch-based approaches like Style-Swap and CNNMRF, which is beyond the capability of SOTA diversified methods, e.g., DFP.

Baseline	Method	D_{pixel}	D_{LPIPS}	Efficiency
MTS	Original	0.080	0.175	-
ITN	Original	0.077	0.163	-
WCT	Original	0.000	0.000	3.421s
	+ DFP	0.162	0.431	4.091s
	+ DivSwapper	**0.204**	**0.485**	**3.566s**
Avatar-Net	Original	0.000	0.000	3.920s
	+ DFP	0.102	0.264	4.268s
	+ DivSwapper	**0.128**	**0.320**	**3.932s**
Style-Swap	Original	0.000	0.000	10.571s
	+ DivSwapper	**0.065**	**0.234**	**10.582s**
[1]CNNMRF	Original	0.084	0.257	118.44s
	+ DivSwapper	**0.142**	**0.378**	**140.91s**

[1] Due to the limitation of GPU memory, we only test the images of size 448×448px for CNNMRF.

Table 1: Quantitative comparisons. The efficiency is tested on images of size 512×512px and a 6GB Nvidia 1060 GPU.

colors changed on the skies and buildings in rows (d,f) (*best compared with the difference heat maps in the last column*). Moreover, as shown in Fig. 1 and 7, our DivSwapper can also diversify the pure patch-based methods like Style-Swap and CNNMRF, which is beyond the capability of DFP. It is worth noting that the patterns varied significantly in our results in Fig. 7 generally correspond to the style regions with higher activation values in the activation heat maps, e.g., the beige walls in the top and the blue and red edges in the bottom. It verifies that our DivSwapper indeed helps vary more significant style patches with higher activation values. *We also validate its effectiveness on AdaIN [Huang and Belongie, 2017] and SANet [Park and Lee, 2019] in SM.*

Quantitative Comparison. The quantitative results are shown in Tab. 1. Consistent with Fig. 6, MTS and ITN obtain low diversity scores in both Pixel and LPIPS distance. Integrated into the same baselines (i.e., WCT and Avatar-Net), our DivSwapper is clearly superior to DFP in both diversity and efficiency (DFP involves some slow CPU-based SVD operations to obtain orthogonal noise matrix). In addition, our DivSwapper can also diversify Style-Swap and help improve the diversity of CNNMRF. Note that due to the use of noise initialization and iterative optimization process, CNNMRF has produced some varied results and the extra time increased by our DivSwapper is more than other baselines.

Baseline	Original	+ DFP	+ Our DivSwapper
WCT	27.24	33.96	**38.80**
Avatar-Net	28.81	31.78	**39.41**

Table 2: Percentage (%) of the votes in the user study.

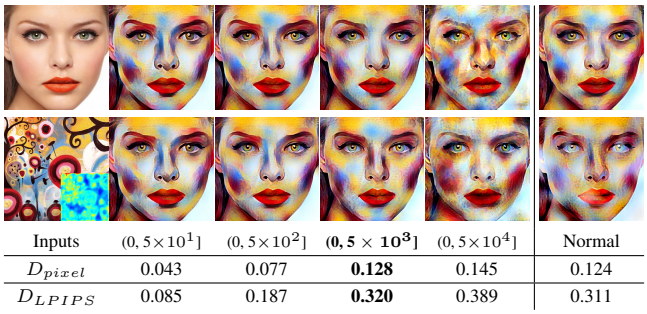

Figure 8: Effects of different sampling ranges (2^{nd} to 5^{th} columns) and distributions (last column) of the random deviations σ. Our DivSwapper is integrated into Avatar-Net [Sheng *et al.*, 2018].

Quality Comparison. As style transfer is highly subjective, we conduct a user study to evaluate how users may prefer the outputs of our diversified methods over the deterministic ones and those of other SOTA diversified methods (i.e., DFP). WCT and Avatar-Net are adopted as the baselines. Twenty users unconnected with the project are recruited. For each baseline, we give each user 50 groups of images (each group contains the input content and style images, and three randomly shuffled outputs, i.e., one original output of the baseline method, one random output of baseline + DFP, and one random output of baseline + our DivSwapper) and ask him/her to select the favorite output. The statistics in Tab. 2 show that both DFP and our DivSwapper can help users obtain preferred (higher quality) results compared with baselines, and our method also achieves higher quality than DFP.

4.3 Ablation Study

Graceful Control between Diversity and Quality. Our DivSwapper can provide graceful control between diversity and quality by sampling the deviations σ from different ranges. As shown in Fig. 8, with the increase of sampling range, the generated results consistently gain more diversity (which can also be validated by the metrics below), but a too-large sampling range may reduce the quality (e.g., the 5^{th} column). When proper range (e.g., $(0, 5 \times 10^3]$) is applied, we can obtain the sweet spot of the two: the results exhibit considerable diversity and also maintain the original quality and some inherent characteristics. For different baselines, the proper range of σ can be easily determined via only a few trials and errors, and our experiments verify that *these constant range values can work stably on different content and style inputs*.

Effect of Sampling Distribution. We also try other sampling distributions instead of the default uniform one. As shown in the last column of Fig. 8, sampling σ from a normal distribution could achieve similar performance (e.g., the top image and the diversity scores), but the results may be er-

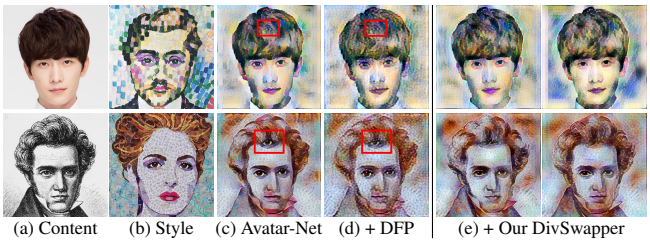

Figure 9: Results on semantic-level style transfer, e.g., portrait-to-portrait. The diverse results generated by using our DivSwapper can maintain the semantic-level stylization while also alleviating the undesirable artifacts (e.g., the eye patterns in the red box areas) caused by the restricted patch matching.

ratic and sometimes produce unwanted effects (e.g., the hazy blocks in the bottom image). This problem may be caused by the concentration property of the normal distribution. However, it does not occur when using a uniform distribution.

Semantic-level Style Transfer. Though our primary motivation is to improve the diversity in *artistic style transfer*, for semantic-level style transfer, the proposed method can also produce diverse results while maintaining the original quality and the semantic correspondence. As can be seen in our generated results in Fig. 9 (e), although the patterns in each semantic area vary significantly (e.g., the backgrounds and hairs), the main semantic-level stylization is well preserved (e.g., the details on portraits). It is because the SSN used in our DivSwapper is still constrained by the original NCC, and the variations are gracefully controlled by σ, as analyzed in Sec. 3.2. Moreover, our DivSwapper can also help alleviate the inherent flaws (e.g., undesirable artifacts) caused by the original restricted patch matching [Zhang *et al.*, 2019], which is our new merit against SOTA methods, e.g., DFP in column (d). It also further justifies that our method can help users obtain diverse results with higher quality.

5 Concluding Remarks

In this work, we explore the challenging problem of diversified patch-based style transfer and introduce a universal and efficient module, i.e., *DivSwapper*, to resolve it. Our DivSwapper is plug-and-play and can be easily integrated into existing patch-based and Gram-based methods to generate diverse results for arbitrary styles. Theoretical analyses and extensive experiments demonstrate the effectiveness of our method, and compared with SOTA algorithms, it shows superiority in diversity, quality, and efficiency. We hope our analyses and investigated method can help readers better understand the crux of patch-based methods and inspire future works in style transfer and many other similar fields.

Acknowledgements

This work was supported in part by the projects No. 2021YFF0900604, 19ZDA197, LY21F020005, 2021009, 2019011, Zhejiang Elite Program project: research and application of media fusion digital intelligence service platform based on multimodal data, MOE Frontier Science Center for Brain Science & Brain-Machine Integration (Zhejiang University), National Natural Science Foundation of China (62172365), and Key Scientific Research Base for Digital Conservation of Cave Temples (Zhejiang University), State Administration for Cultural Heritage.

References

[Aberman *et al.*, 2018] Kfir Aberman, Jing Liao, Mingyi Shi, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Neural best-buddies: Sparse cross-domain correspondence. *TOG*, 37(4):1–14, 2018.

[An *et al.*, 2021] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *CVPR*, pages 862–871, 2021.

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

[Chandran *et al.*, 2021] Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, Markus Gross, and Derek Bradley. Adaptive convolutions for structure-aware style transfer. In *CVPR*, pages 7972–7981, 2021.

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

[Chen *et al.*, 2017] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *CVPR*, pages 1897–1906, 2017.

[Chen *et al.*, 2021a] Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Artistic style transfer with internal-external learning and contrastive learning. In *NeurIPS*, 2021.

[Chen *et al.*, 2021b] Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Dualast: Dual style-learning networks for artistic style transfer. In *CVPR*, pages 872–881, 2021.

[Cheng *et al.*, 2021] Jiaxin Cheng, Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Prem Natarajan. Style-aware normalized loss for improving arbitrary style transfer. In *CVPR*, pages 134–143, 2021.

[Deng *et al.*, 2020] Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. Arbitrary style transfer via multi-adaptation network. In *ACM MM*, pages 2719–2727, 2020.

[Gatys *et al.*, 2016] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016.

[Gu *et al.*, 2018] Shuyang Gu, Congliang Chen, Jing Liao, and Lu Yuan. Arbitrary style transfer with deep feature reshuffle. In *CVPR*, pages 8222–8231, 2018.

[Huang and Belongie, 2017] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017.

[Jing *et al.*, 2019] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *TVCG*, 26(11):3365–3385, 2019.

[Jing *et al.*, 2020] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. Dynamic instance normalization for arbitrary style transfer. In *AAAI*, volume 34, pages 4369–4376, 2020.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[Kolkin *et al.*, 2019] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *CVPR*, pages 10051–10060, 2019.

[Li and Wand, 2016a] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *CVPR*, pages 2479–2486, 2016.

[Li and Wand, 2016b] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, pages 702–716. Springer, 2016.

[Li *et al.*, 2017a] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *IJCAI*, pages 2230–2236, 2017.

[Li *et al.*, 2017b] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Diversified texture synthesis with feed-forward networks. In *CVPR*, 2017.

[Li *et al.*, 2017c] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *NeurIPS*, pages 386–396, 2017.

[Li *et al.*, 2019] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *CVPR*, pages 3809–3817, 2019.

[Liao *et al.*, 2017] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *TOG*, 2017.

[Lin *et al.*, 2021] Tianwei Lin, Zhuoqi Ma, Fu Li, Dongliang He, Xin Li, Errui Ding, Nannan Wang, Jie Li, and Xinbo Gao. Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer. In *CVPR*, pages 5141–5150, 2021.

[Liu *et al.*, 2021] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *ICCV*, pages 6649–6658, 2021.

[Lu *et al.*, 2019] Ming Lu, Hao Zhao, Anbang Yao, Yurong Chen, Feng Xu, and Li Zhang. A closed-form solution to universal style transfer. In *ICCV*, pages 5952–5961, 2019.

[Park and Lee, 2019] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *CVPR*, pages 5880–5888, 2019.

[Sheng *et al.*, 2018] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *CVPR*, pages 8242–8250, 2018.

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

[Ulyanov *et al.*, 2016] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, pages 1349–1357, 2016.

[Ulyanov *et al.*, 2017] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*, pages 6924–6932, 2017.

[Wang *et al.*, 2020a] Zhizhong Wang, Lei Zhao, Haibo Chen, Lihong Qiu, Qihang Mo, Sihuan Lin, Wei Xing, and Dongming Lu. Diversified arbitrary style transfer via deep feature perturbation. In *CVPR*, pages 7789–7798, 2020.

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

[Wang *et al.*, 2021] Zhizhong Wang, Lei Zhao, Haibo Chen, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Evaluate and improve the quality of neural style transfer. *CVIU*, 207:103203, 2021.

[Wang *et al.*, 2022] Zhizhong Wang, Lei Zhao, Haibo Chen, Ailin Li, Zhiwen Zuo, Wei Xing, and Dongming Lu. Texture reformer: Towards fast and universal interactive texture transfer. In *AAAI*, 2022.

[Yao *et al.*, 2019] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. Attention-aware multi-stroke style transfer. In *CVPR*, pages 1467–1475, 2019.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[Zhang *et al.*, 2019] Yulun Zhang, Chen Fang, Yilin Wang, Zhaowen Wang, Zhe Lin, Yun Fu, and Jimei Yang. Multimodal style transfer via graph cuts. In *ICCV*, pages 5943–5951, 2019.