Diverse Image Style Transfer via Invertible Cross-Space Mapping

Haibo Chen, Lei Zhao*, Huiming Zhang, Zhizhong Wang
Zhiwen Zuo, Ailin Li, Wei Xing*, Dongming Lu

College of Computer Science and Technology, Zhejiang University
{cshbchen, cszhl, qinglanwuji, endywon, zzwcs, liailin, wxing, ldm}@zju.edu.cn

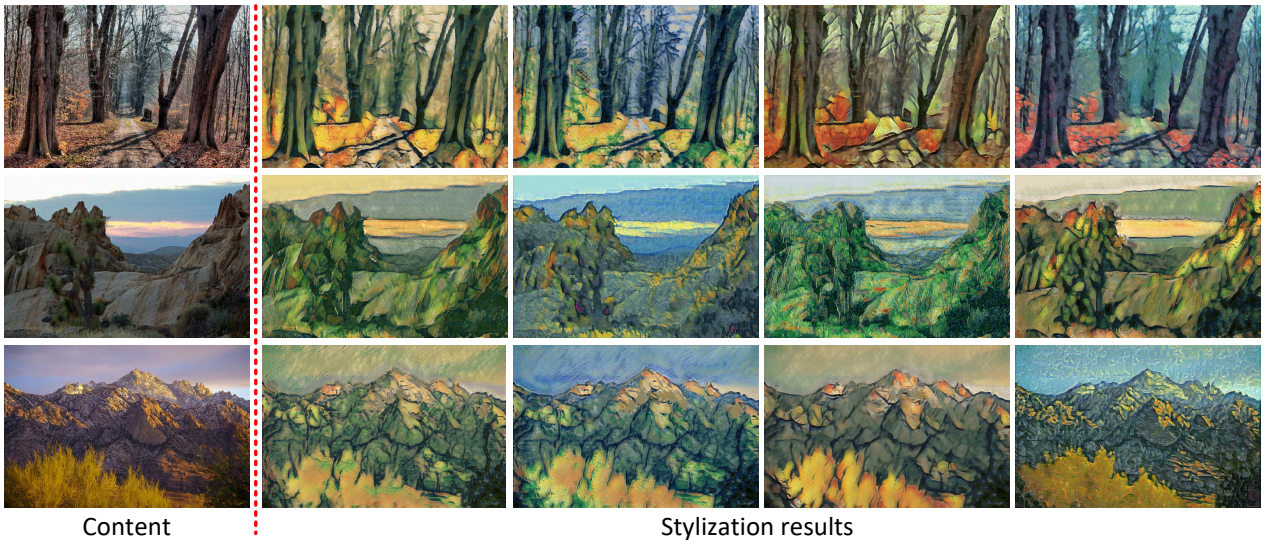

Figure 1: Stylization examples generated by our proposed DIST. The first column shows the content images. The other four columns show the diverse stylization results based on Paul Cezanne's style.

Abstract

Image style transfer aims to transfer the styles of artworks onto arbitrary photographs to create novel artistic images. Although style transfer is inherently an underdetermined problem, existing approaches usually assume a deterministic solution, thus failing to capture the full distribution of possible outputs. To address this limitation, we propose a Diverse Image Style Transfer (DIST) framework which achieves significant diversity by enforcing an invertible cross-space mapping. Specifically, the framework consists of three branches: disentanglement branch, inverse branch, and stylization branch. Among them, the disentanglement branch factorizes artworks into content space and style space; the inverse branch encourages the invertible mapping between the latent space of input noise vectors and the style space of generated artistic images; the stylization branch renders the input content image with the style of an artist. Armed with these three branches, our approach is able to synthesize significantly diverse stylized images without loss of quality. We conduct extensive experiments and comparisons to evaluate our approach qualitatively and quantitatively. The experimental results demonstrate the effectiveness of our method.

1. Introduction

An exquisite artwork can take a diligent artist days or even months to create, which is labor-intensive and time-consuming. Motivated by this, a series of recent approaches studied the problem of repainting an existing photograph with the style of an artist using either a single artwork or a collection of artworks. These approaches are known as style transfer. Armed with style transfer techniques, anyone could create artistic images.

* Corresponding author

How to represent the content and style of an image is the key challenge of style transfer. Recently, the seminal work of Gatys *et al*. [7] firstly proposed to extract content and style features from an image using pre-trained Deep Convolutional Neural Networks (DCNNs). By separating and recombining contents and styles of arbitrary images, novel artworks can be created. This work showed the enormous potential of CNNs in style transfer and created a surge of interest in this field. Based on this work, a series of subsequent methods have been proposed to achieve better performance in many aspects, including efficiency [13, 21, 34], quality [20, 35, 40, 43, 39, 4], and generalization [6, 5, 10, 24, 30, 27, 22]. However, diversity, as another important aspect, has received relatively less attention.

As the saying goes, *"There are a thousand Hamlets in a thousand people's eyes"*. Similarly, different people have different understanding and interpretation of the style of an artwork. There is no uniform and quantitative definition of the artistic style of an image. Therefore, the stylization results should be diverse rather than unique, so that the preferences of different people can be satisfied. To put it another way, style transfer is an underdetermined problem, where a large number of solutions can be found. Unfortunately, existing style transfer methods usually assume a deterministic solution. As a result, they fail to capture the full distribution of possible outputs.

A straightforward approach to handle diversity in style transfer is to take random noise vectors along with content images as inputs, *i.e*., utilizing the variability of the input noise vectors to produce diverse stylization results. However, the network tends to pay more attention to the high-dimensional and structured content images and ignores the noise vectors, leading to deterministic output. To ensure that the variability in the latent space can be passed into the image space, Ulyanov *et al*. [35] enforced the dissimilarity among generated images by enlarging their distance in the pixel space. Similarly, Li *et al*. [23] introduced a diversity loss that penalized the feature similarities of different samples in a mini-batch. Although these methods can achieve diversity to some extent, they have obvious limitations. First, forcibly enlarging the distance among outputs may cause the results to deviate from the local optimum, resulting in the degradation of image quality. Second, to avoid introducing too many artifacts to the generated images, the weight of the diversity loss is generally set to a small value. Consequently, the diversity of the stylization results is relatively limited. Third, diversity is more than the pixel distance or feature distance among generated images, which contains richer and more complex connotation. Most recently, Wang *et al*. [37] achieved better diversity by using an orthogonal noise matrix to perturb the image feature maps while keeping the original style information unchanged. However, this approach is apt to generate distorted results, providing insufficient visual quality. Therefore, the problem of diverse style transfer remains an open challenge.

In this paper, we propose a Diverse Image Style Transfer (DIST) framework which achieves significant diversity without loss of quality by enforcing an invertible cross-space mapping. Specifically, the framework takes random noise vectors along with everyday photographs as its inputs, where the former are responsible for style variations and the latter determine the main contents. However, according to above analyses, we can learn that the noise vectors are prone to be ignored in the network. Our proposed DIST framework tackles this problem through three branches: disentanglement branch, inverse branch, and stylization branch.

The disentanglement branch factorizes artworks into content space and style space. *The inverse branch* encourages the invertible mapping between the latent space of input noise vectors and the style space of generated artistic images, which is inspired by [32]. But different from [32], we invert the style information rather than the whole generated image to the input noise vector, since the input noise vector mainly influences the style of the generated image. *The stylization branch* renders the input content image with the style of an artist. Equipped with these three branches, DIST is able to synthesize significantly diverse stylized images without loss of quality, as shown in Figure 1.

Overall, the contributions can be summarized as follows:

- We propose a novel style transfer framework which achieves significant diversity by learning the one-to-one mapping between latent space and style space.
- Different from existing style transfer methods [35, 23, 37] that obtain diversity with serious degradation of quality, our approach can produce both high-quality and diverse stylization results.
- Our approach provides a new way to disentangle the style and content of an image.
- We demonstrate the effectiveness and superiority of our approach by extensive comparison with several state-of-the-art style transfer methods.

2. Related Work

Style Transfer. Style transfer aims at synthesizing new images with artistic styles by repainting an existing photograph utilizing the style information extracted from real artworks. Gatys *et al*. [7] first proposed to separate and recombine arbitrary images' contents and styles, which are captured from a pre-trained VGG-19 network [31], to generate novel artistic images. This method is capable of producing striking stylization results, but is prohibitively slow due to the iterative optimization process. To enable faster stylization, [13, 21, 34] proposed to utilize feed-forward networks

to efficiently synthesize a stylized image. However, these methods, while enjoying the inference efficiency, are often limited by compromised visual quality. Motivated by this, a lot of methods [20, 36, 40, 43, 38] have been proposed to enhance the quality of generated images from different aspects. Another line of work focused on improving the generalization of style transfer networks, and developed a number of arbitrary style transfer methods [5, 10, 24, 30, 27, 12].

Above style transfer methods extract style representations from a single artwork. Sanakoyeu *et al*. [29] pointed out that it is insufficient to only use a single artwork, because it might not represent the full scope of an artistic style. Therefore, [29] proposed to learn style from a collection of related artworks, vastly boosting the visual quality. [17, 18, 33] are three follow-up works. [17] could capture subtle variations of style while also being able to distinguish different styles and disentangle content and style. [18] proposed a content-transformation block to alter an object in a content- and style-specific manner. Svoboda *et al*. [33] achieved zero-shot style transfer with a novel two-stage peer-regularization layer. *In this paper, we follow this line of work and focus on the diversity problem that is ignored by these methods.*

Diverse Image Generation. Currently there are many generative models that are able to generate diverse output images, among which Generative Adversarial Networks (GANs) [8, 28, 2, 26, 41, 3] may be the most well-know one. The core idea of GANs lies in the adversarial loss that enforces the distribution of generated images to match the real data distribution. However, GANs often suffer from mode collapse. To resolve this problem, Srivastava *et al*. [32] proposed to encourage the one-to-one relationship between the input noise vector and the generated image, thereby significantly improving the diversity of generated images. Kazemi *et al*. [16] further introduced SC-GAN for content and style disentangled representation learning. In detail, they enforced the correspondence between the content/style code of the input noise vector and the content/style information of the generated image. Consequently, by fixing the content portion of the input, they can generate a specific scene in a variety of styles. The other portion is analogous.

Above methods all aim at a noise-to-image generation problem, while style transfer is an image-to-image translation problem. It is much harder to achieve diversity in the image-to-image translation scenario, since the noise vectors (which are responsible for diversity) are prone to be ignored when high-dimensional and structured images are also taken as inputs along with the noise vectors [25, 1].

Similar to style transfer, image domain translation is also an image-to-image translation problem, where the goal is to learn the mapping between different yet similar visual domains, for example, horses↔zebras. To achieve diversity, Zhu *et al*. [46] proposed a BicycleGAN that can model continuous and multimodal distributions, which shares similar spirits with [32]. Nevertheless, the method is only applicable to problems with paired training data. Motivated by this, [11, 19] proposed diverse unsupervised image domain translation methods, which are based on the assumption: the image representation can be decomposed into a domain-invariant content space that captures shared information across domains, and a domain-specific style space that can model the diverse variations given the same content. Different from image domain translation, in style transfer, the content image and style image usually contain totally different contents, which suggests that above methods are inapplicable to style transfer. By far, only a few attempts have been made to enforce diversity in style transfer. [35, 23] proposed to maximize the distance among stylized images. [37] employed a deep feature perturbation (DFP) operation to perturb the deep image feature maps. Although these methods can achieve diversity to some extent, they sacrificed the quality of generated images. In this paper, we propose a novel style transfer approach which achieves better diversity without sacrificing visual quality.

3. Approach

Inspired by [29, 17, 18, 33], we learn artistic style not from a single artwork but from a collection of related artworks. Formally, our task can be described as follows: given a collection of photos $x \sim X$ and a collection of artworks $y \sim Y$ (the contents of X and Y can be totally different), we aim to learn a style transformation $\mathcal{G} : X \rightarrow Y$ with significant diversity. To achieve this goal, we propose a DIST framework consisting of three branches: stylization branch, disentanglement branch, and inverse branch. In this section, we introduce the three branches in details.

3.1. Stylization Branch

The stylization branch aims to repaint $x \sim X$ with the style of $y \sim Y$. To this end, we enable $\mathcal{G}$ to approximate the distribution of Y by employing a discriminator $\mathcal{D}$ to train against $\mathcal{G}$: $\mathcal{G}$ tries to generate images that resembles the images in Y, while $\mathcal{D}$ tries to distinguish the stylized images from the real ones. Joint training of these two networks leads to a generator that is able to produce desired stylizations. This process can be formulated as follows (note that for $\mathcal{G}$, we adopt an encoder-decoder architecture consisting of an encoder E_c and a decoder D) :

$$\mathcal{L}_{adv} := \mathop{\mathbb{E}}_{y\sim Y}[log(\mathcal{D}(y))] + \mathop{\mathbb{E}}_{x\sim X, z\sim p(z)} [log(1 - \mathcal{D}(D(E_c(x), z)))] \quad (1)$$

where $z \in \mathbb{R}^{d_z}$ is a random noise vector and $p(z)$ is the standard normal distribution $\mathcal{N}(0, I)$. We leverage its variability to encourage diversity in generated images.

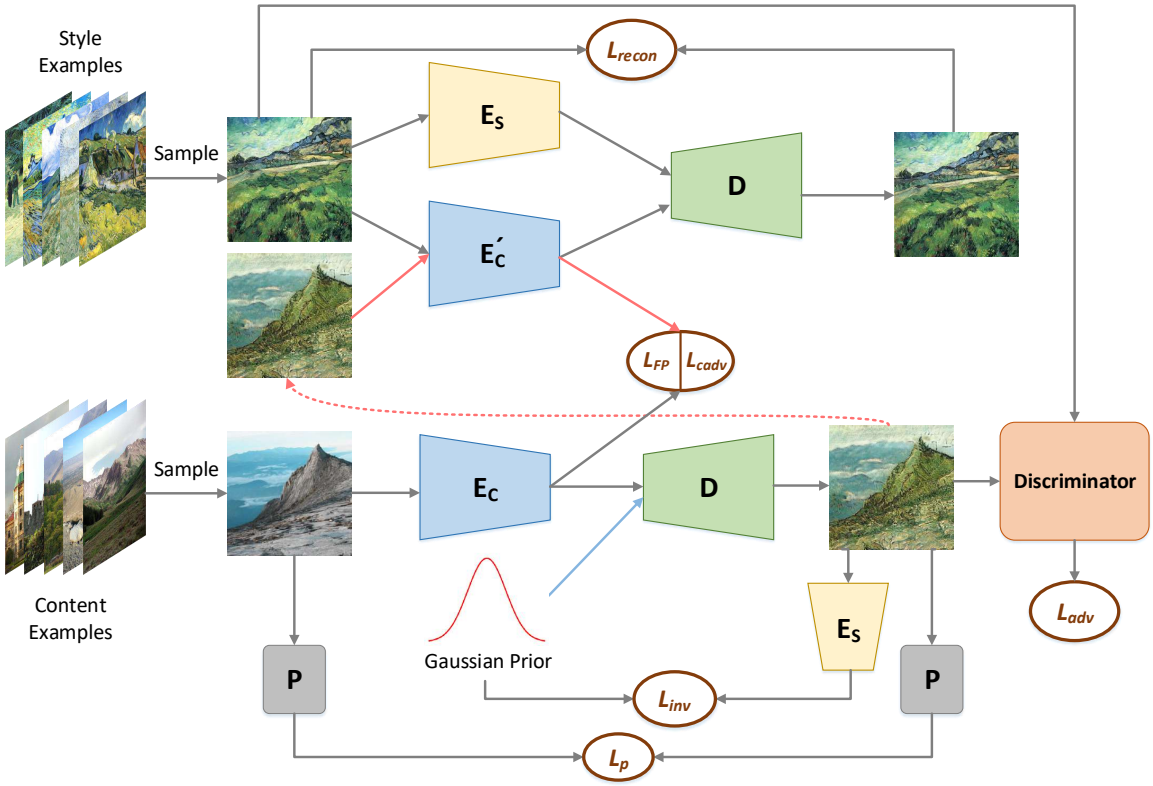

Figure 2: An overview of our method. (a) With the adversarial loss $\mathcal{L}_{adv}$ and content structure loss $\mathcal{L}_p$ (Section 3.1), we are able to transfer artistic styles onto content images. (b) With the content feature loss $\mathcal{L}_{FP}$, content feature adversarial loss $\mathcal{L}_{cadv}$, and artwork reconstruction loss $\mathcal{L}_{recon}$ (Section 3.2), we obtain an encoder E_s which can extract the style information from stylized images. (c) With the inverse loss $\mathcal{L}_{inv}$ (Section 3.3), we encourage the bijection mapping between the style space of stylized images and the latent space of input noise vectors, leading to significant diversity.

Only using above adversarial loss cannot preserve the content information of x in the generated image, which does not meet the requirements of style transfer. The simplest solution is to utilize a pixel-wise loss between the content image $x \sim X$ and stylized image $D(E_c(x), z)$. However, this loss is too strict and harms the quality of the stylized image. Therefore, we soften the constraint: instead of directly calculating the distance between original images, we first input them into an average pooling layer P and then calculate the distance between them. We express this content structure loss as:

$$\mathcal{L}_p := \mathop{\mathbb{E}}_{x\sim X, z\sim p(z)} [||P(D(E_c(x), z)) - P(x)||_2^2] \quad (2)$$

Compared with the pixel-wise loss which requires the content image and the stylized image to be exactly the same, $\mathcal{L}_p$ measures their difference in a more coarse-grained manner and only requires them to be similar in general content structures, more consistent with the goal of style transfer.

Although the stylization branch is sufficient to obtain remarkable stylized images, it can only produce a deterministic stylized image without diversity, because the network tends to ignore the random noise vector z.

3.2. Disentanglement Branch

[32] alleviated the mode collapse issue in GANs by enforcing the bijection mapping between the input noise vectors and generated images. Different from [32], which only takes noise vectors as inputs, our model takes noise vectors along with content images as inputs, where the former are responsible for style variations and the latter determine the main contents. Therefore, in the inverse process, instead of inverting the whole generated image to the input noise vector like [32] does, we invert the style information of the stylized image to the input noise vector (details in Section 3.3). To be specific, we utilize a style encoder to extract the style information from the stylized image, and enforce the consistency between the style encoder's output and the input noise vector. The main problem now is how to obtain such a style encoder. We resolve this problem through the disentanglement branch.

First, the disentanglement branch employs an encoder $E_c^{'}$ which takes the stylized image $D(E_c(x), z)$ as input. Given that the content image and stylized image share the same content and differ greatly in style, if we encourage the similarity between the output of E_c (whose input is the content image) and that of $E_c^{'}$ (whose input is the stylized image), E_c and $E_c^{'}$ shall extract the shared content information and neglect the specific style information. Notice that E_c and $E_c^{'}$ are two independent networks and do not share weights. This is because there are some differences when extracting photographs' contents and artworks' contents. We define the corresponding content feature loss as,

$$\mathcal{L}_{FP} := \mathop{\mathbb{E}}_{x\sim X, z\sim p(z)}[||E_c^{'}(D(E_c(x), z)) - E_c(x)||_2^2] \quad (3)$$

However, $\mathcal{L}_{FP}$ may encourage E_c and $E_c^{'}$ to output feature maps in which the value of each element is pretty small (*i.e.*, $\| \ E_c(x) \ \| \to 0, \| \ E_c^{'}(D(E_c(x), z)) \ \| \to 0$). In such a circumstance, although $\mathcal{L}_{FP}$ is minimized, the similarity between $E_c(x)$ and $E_c^{'}(D(E_c(x), z))$ is not increased. To alleviate this problem, we employ a feature discriminator $\mathcal{D}_f$ and introduce a content feature adversarial loss,

$$\begin{aligned}\mathcal{L}_{cadv} := \mathop{\mathbb{E}}_{x\sim X, z\sim p(z)}[log(\mathcal{D}_f(E_c(x))) + \\ log(1 - \mathcal{D}_f(E_c^{'}(D(E_c(x), z))))]\end{aligned} \quad (4)$$

$\mathcal{L}_{cadv}$ measures the distribution deviation, less sensitive to the value of its input in comparison with $\mathcal{L}_{FP}$. In addition, $\mathcal{L}_{cadv}$ together with $\mathcal{L}_{FP}$ can promote the similarity in two dimensions, further improving the performance.

Then the disentanglement branch adopts another encoder E_s together with the content encoder $E_c^{'}$ and the decoder D to reconstruct the artistic image. Since $E_c^{'}$ is constrained to extract the content information, E_s has to extract the style information to reconstruct the artistic image. Therefore, we get our desired style encoder E_s. We formulate the reconstruction loss as,

$$\mathcal{L}_{recon} := \mathop{\mathbb{E}}_{y\sim Y}[||D(E_c^{'}(y), E_s(y)) - y||_1] \quad (5)$$

3.3. Inverse Branch

Armed with the style encoder E_s, we can access the style space of artistic images. To achieve diversity, the inverse branch enforces the one-to-one mapping between latent space and style space by employing the inverse loss,

$$\mathcal{L}_{inv} := \mathop{\mathbb{E}}_{x\sim X, z\sim p(z)}[||E_s(D(E_c(x), z)) - z||_1] \quad (6)$$

The inverse loss ensures that the style information of the generated image $D(E_c(x), z)$ can be inverted to the corresponding noise vector z, which implies that $D(E_c(x), z)$ retains the influence and variability of z. In this way, we can get diverse stylization results by randomly sampling different z from the standard normal distribution $\mathcal{N}(0, I)$.

3.4. Final Objective and Network Architectures

Figure 2 illustrates the full pipeline of our approach. We summarize all aforementioned losses and obtain the compound loss,

$$\begin{aligned}\mathcal{L}_{total} := \lambda_{adv}\mathcal{L}_{adv} + \lambda_p\mathcal{L}_p + \lambda_{fp}\mathcal{L}_{FP} + \\ \lambda_{cadv}\mathcal{L}_{cadv} + \lambda_{recon}\mathcal{L}_{recon} + \lambda_{inv}\mathcal{L}_{inv}\end{aligned} \quad (7)$$

where the hyper-parameters λ_{adv}, λ_p, λ_{fp}, λ_{cadv}, λ_{recon}, and λ_{inv} control the importance of each term. We use the compound loss as the final objective to train our model.

Network Architectures. We build on the recent AST backbone [29], and extend it with our proposed changes to produce diverse stylization results. Specifically, the content encoder E_c and $E_c^{'}$ have the same architecture and are composed of five convolution layers. The style encoder E_s includes five convolution layers, a global average pooling layer, and a fully connected (FC) layer. Similar to [15], our decoder D has two branches. One branch takes the content image x as input, containing nine residual blocks [9], four upsampling blocks, and one convolution layer. Another branch takes the noise vector z as input (*notice that at inference time, we can take either z or the style code $E_s(y)$ extracted from a reference image y as its input*), containing one FC layer to produce a set of affine parameters γ, β. Then the two branches are combined through AdaIN [10],

$$AdaIN(a, \gamma, \beta) := \gamma\left(\frac{a - \mu(a)}{\sigma(a)}\right) + \beta \quad (8)$$

where a is the activation of the previous convolutional layer in branch one, μ and σ are channel-wise mean and standard deviation, respectively. The image discriminator $\mathcal{D}$ is a fully convolutional network with seven convolution layers. The feature discriminator $\mathcal{D}_f$ consists of three convolution layers and one FC layer. As for P, it is an average pooling layer. The loss weights are set to λ_{adv} = 2, λ_p = 150, λ_{fp} = 100, λ_{cadv} = 10, λ_{recon} = 200, and λ_{inv} = 600. We use the Adam optimizer with a learning rate of 0.0002.

4. Experiments

We conduct extensive experiments and comparisons to evaluate our proposed method. First, in Section 4.1, we show the diverse artworks generated by our model and perform qualitative comparisons. Next, we provide quantitative results in Section 4.2. Finally, in Section 4.3, we ablate single components of our model to show their importance.

Dataset. Like [29, 17, 18, 33], we take Places365 [45] as the content dataset and Wikiart [14] as the style

Figure 3: Stylization examples generated by DIST. The first row shows the artworks of different artists. The second row shows the content images. The other three rows show the diverse stylized images generated by our model.

dataset (concretely, we collect hundreds of artworks for each artist from WikiArt and train a separate model for him/her). Training images were randomly cropped and resized to 768×768 resolutions.

Baselines. We take the following methods that can produce diversity as our baselines: Gatys *et al*. [7], Li *et al*. [23], Ulyanov *et al*. [35], DFP [37], and MUNIT [11]. Apart from above methods, we also compare with AST [29] and Svoboda *et al*. [33] to make the experiments more sufficient. Note that we use their officially released codes and default settings of hyper-parameters for experiments.

4.1. Qualitative Comparisons

In this section, we present images generated by our method to confirm the quantitative results in terms of diversity and quality. Figure 3 shows our stylization results based on different artists' styles. We can see that for each artist's style, our model produces significantly diverse artistic images with remarkable visual quality.

In Figure 4, we show the qualitative comparison results between the proposed DIST and the seven baselines mentioned above. We observe that AST [29] and Svoboda *et al*. [33] fail to generate diverse outputs. Gatys *et al*. [7], Li *et al*. [23], and Ulyanov *et al*. [35] only produce slight variations, which are hard to notice. DFP [37] achieves noticeable diversity but introduces many distortions in the stylized image, failing to preserve the main content structures. MUNIT [11] yields highly diverse yet poor-quality stylizations. As can be seen from the zoom-in part in Figure 4, MUNIT [11] only changes the color of the content image and does not learn any texture patterns, resulting in unsatisfying results. This is because MUNIT [11] is built on the assumption: images in different domains have different style spaces but share a common content space, which implies that it can only perform image translation between visually similar domains (for example, day scene↔night scene). In contrast, DIST does not require the content image and style image to be similar in content. The results in Figure 3 and 4 verify the effectiveness and superiority of our method. Additional results are provided in the **supplementary material.**

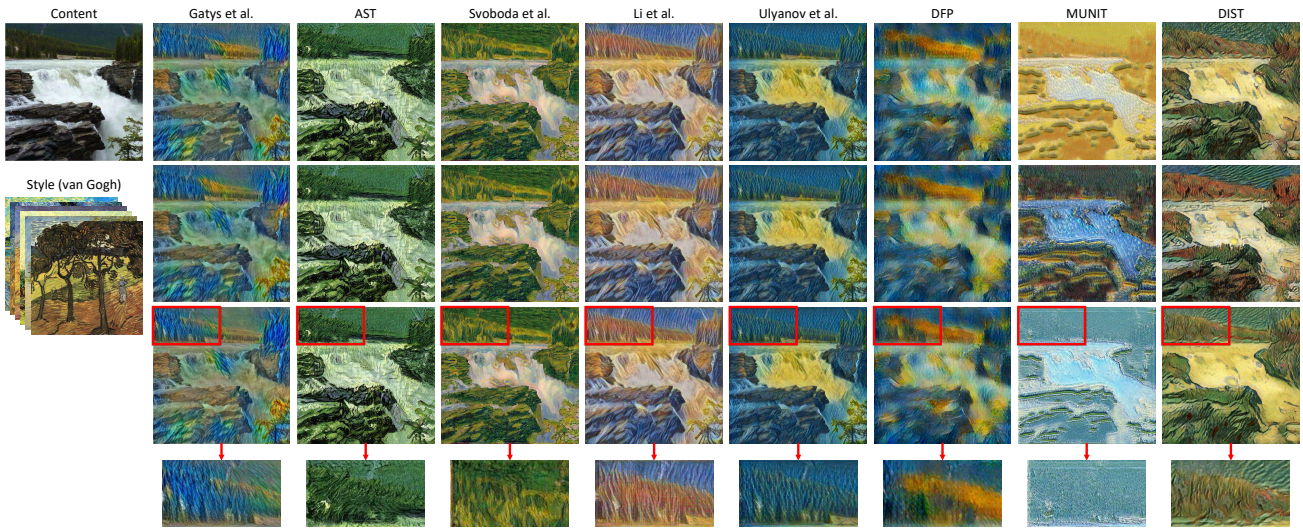

Figure 4: Qualitative comparisons. The first column shows the content image and van Gogh's artworks. The other columns show the stylized images generated by different methods.

Table 1: The deception rate, user study, and LPIPS distance for different methods. The higher the better.

	Gatys *et al*. [7]	AST [29]	Svoboda *et al*. [33]	Li *et al*. [23]	Ulyanov *et al*. [35]	DFP [37]	MUNIT [11]	DIST
Deception Rate	0.206	0.454	0.278	0.072	0.079	0.027	0.121	**0.525**
User Study	0.089	0.316	0.242	0.010	0.012	0.006	0.004	**0.321**
LPIPS Distance	0.256	0.000	0.000	0.175	0.163	0.431	**0.538**	0.464

4.2. Quantitative Comparisons

In this section, we assess our model with some evaluation metrics. Specifically, we adopt deception rate [29] and user study [27, 40, 4, 44] to measure quality, and employ LPIPS (*Learned Perceptual Image Patch Similarity*) distance [42] to measure diversity.

Deception rate. This is a quantitative metric proposed by Sanakoyeu *et al*. [29]. In particular, a VGG-16 network [31] was pre-trained to classify artists on Wikiart [14]. The deception rate is then calculated as the fraction of generated images which were classified by the network as the artworks of an artist for which the stylization was produced. We report the deception rate in Table 1 in the second row, where we can see that our approach performs the best while DFP [37] performs the worst.

User study. We also perform human evaluation studies to compare the performance of DIST with other methods. Given a content image, we stylize it with different methods and show the stylization results alongside the content image to participants. We then ask these participants to choose the stylized image which resembles the style of the target artist the most. We show 20 groups of images to 50 participants and finally collect 1000 votes. We report the percentage of votes for each method in the third row of Table 1. We observe that the stylized images generated by DIST are top-rated on average, while MUNIT [11] has the lowest score.

To measure diversity, we use 5 content images and 6 artists' artworks to get 30 different combinations, and for each combination, we require each method to produce 20 outputs. Therefore, we obtain 5700 pairs ($30 \times C_{20}^2 = 5700$) of stylized images generated by each method.

LPIPS distance. LIPIS [42] measures the average feature distances between generated images. The fourth row in Table 1 shows the LPIPS distance of each method. It can be observed that DIST achieves the second-highest score, after MUNIT [11].

In summary, although DFP [37] and MUNIT [11] achieve notable diversity, they perform badly in quality. In contrast, DIST achieves both significant diversity and superior visual quality.

4.3. Ablation Study

According to above analyses, the disentanglement branch and inverse branch are our key to achieving diversity. In this section, we explore the effect of these two branches by ablation studies.

With and without disentanglement branch. To investigate the effect of the disentanglement branch, we evaluate the performance of DIST when this branch is removed. We report the experimental result in Figure 5 (c), where we ob-

Figure 5: Ablation results. (a) The results of AST *(deception rate = 0.454, LPIPS = 0.000)*, (b) DIST *(deception rate = 0.525, LPIPS = 0.464)*, (c) DIST w/o disentanglement branch *(deception rate = 0.112, LPIPS = 0.446)*, and (d) DIST w/o inverse branch *(deception rate = 0.531, LPIPS = 0.034)*. Zoom in for a better view and details.

serve a serious degradation in image quality. The reason could be that it is unreasonable to invert the whole generated image to the input noise vector, since the input noise vector only influences the style of the generated image and has nothing to do with its content.

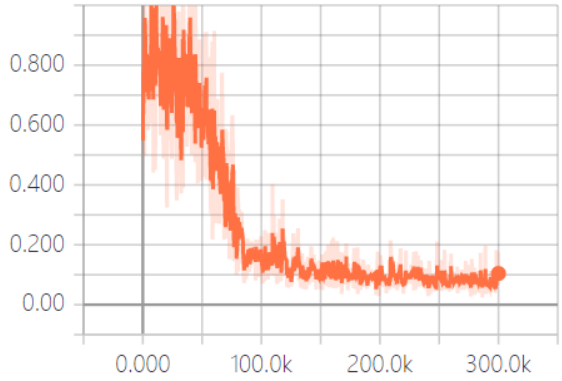

Figure 6: The inverse loss in the training stage.

With and without inverse branch. Here, we train a DIST model that does not involve the inverse branch. As expected, the result in Figure 5 (d) shows that there are few style variations in generated images. This is because the network tends to neglect the input noise vectors. This problem can be solved by employing the proposed inverse branch. As shown in Figure 6, with the inverse branch, the inverse loss is close to 0 at the end of the training stage, suggesting that the DIST model learns an invertible mapping between the latent space and style space. The input noise vectors now can greatly influence the network output.

Note that the noise vector z can be replaced with the style code $E_s(y)$ extracted from a reference image y to produce more controllable stylization results, as shown in Figure 7.

The ablation results indicate that the disentanglement branch and the inverse branch are two essential ingredients of our method. Without these two branches, our method cannot generate diverse and high-quality stylized images.

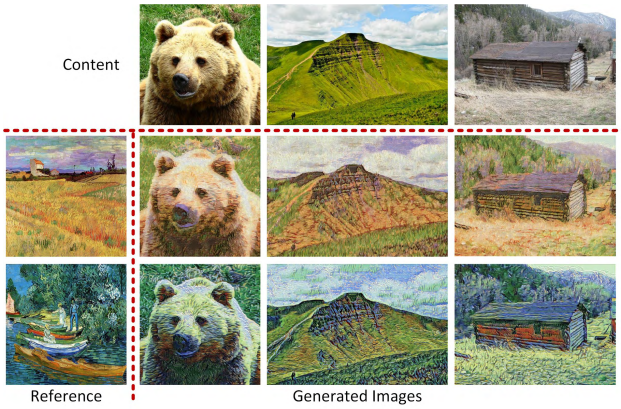

Figure 7: Reference-guided stylization results.

5. Conclusion

In this paper, we propose a Diverse Image Style Transfer (DIST) framework which achieves significant diversity without loss of quality by encouraging the one-to-one mapping between the latent space of input noise vectors and the style space of generated artistic images. The framework consists of three branches, where the stylization branch is responsible for stylizing the content image, and the other two branches (*i.e.*, the disentanglement branch and the inverse branch) are responsible for diversity. Our extensive experimental results demonstrate the effectiveness and superiority of our method. In the future work, we would like to extend our method to other tasks, such as text-to-image synthesis and image inpainting.

Acknowledgments. This work was supported in part by the projects No. 2020YFC1522701, 2020YFC1523101, 19ZDA197, LY21F020005, 2021009, 2020YFC1523201, 2020YFC1523202, 2019C03137, MOE Frontier Science Center for Brain Science & Brain-Machine Integration (Zhejiang University), and Key Scientific Research Base for Digital Conservation of Cave Temples (Zhejiang University), State Administration for Cultural Heritage.

References

[1] Yazeed Alharbi, Neil Smith, and Peter Wonka. Latent filter scaling for multimodal unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1458–1466, 2019. 3

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 3

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 3

[4] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 872–881, 2021. 2, 7

[5] Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, 2016. 2, 3

[6] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 2

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2, 6, 7

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 2, 3, 5

[11] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 3, 6, 7

[12] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. Dynamic instance normalization for arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4369–4376, 2020. 3

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 2

[14] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013. 5, 7

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 5

[16] Hadi Kazemi, Seyed Mehdi Iranmanesh, and Nasser Nasrabadi. Style and content disentanglement in generative adversarial networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 848–856. IEEE, 2019. 3

[17] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4422–4431, 2019. 3, 5

[18] Dmytro Kotovenko, Artsiom Sanakoyeu, Pingchuan Ma, Sabine Lang, and Bjorn Ommer. A content transformation block for image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10032–10041, 2019. 3, 5

[19] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 3

[20] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2479–2486, 2016. 2, 3

[21] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016. 2

[22] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3809–3817, 2019. 2

[23] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Diversified texture synthesis with feed-forward networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3920–3928, 2017. 2, 3, 6, 7

[24] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in neural information processing systems*, pages 386–396, 2017. 2, 3

[25] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1429–1437, 2019. 3

[26] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 3

[27] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 5880–5888, 2019. 2, 3, 7

[28] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 3

[29] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Bjorn Ommer. A style-aware content loss for real-time hd style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 698–714, 2018. 3, 5, 6, 7

[30] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8242–8250, 2018. 2, 3

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 7

[32] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Advances in Neural Information Processing Systems*, pages 3308–3318, 2017. 2, 3, 4

[33] Jan Svoboda, Asha Anoosheh, Christian Osendorfer, and Jonathan Masci. Two-stage peer-regularized feature recombination for arbitrary image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13816–13825, 2020. 3, 5, 6, 7

[34] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *ICML*, volume 1, page 4, 2016. 2

[35] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017. 2, 3, 6, 7

[36] Xin Wang, Geoffrey Oxholm, Da Zhang, and Yuan-Fang Wang. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5239–5247, 2017. 3

[37] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7789–7798, 2020. 2, 3, 6, 7

[38] Zhizhong Wang, Lei Zhao, Haibo Chen, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Evaluate and improve the quality of neural style transfer. *Computer Vision and Image Understanding*, 207:103203, 2021. 3

[39] Zhijie Wu, Chunjin Song, Yang Zhou, Minglun Gong, and Hui Huang. Efanet: Exchangeable feature alignment network for arbitrary style transfer. In *AAAI*, pages 12305–12312, 2020. 2

[40] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. Attention-aware multi-stroke style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1467–1475, 2019. 2, 3, 7

[41] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018. 3

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 7

[43] Yulun Zhang, Chen Fang, Yilin Wang, Zhaowen Wang, Zhe Lin, Yun Fu, and Jimei Yang. Multimodal style transfer via graph cuts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5943–5951, 2019. 2, 3

[44] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5741–5750, 2020. 7

[45] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 5

[46] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017. 3