DualAST: Dual Style-Learning Networks for Artistic Style Transfer

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

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

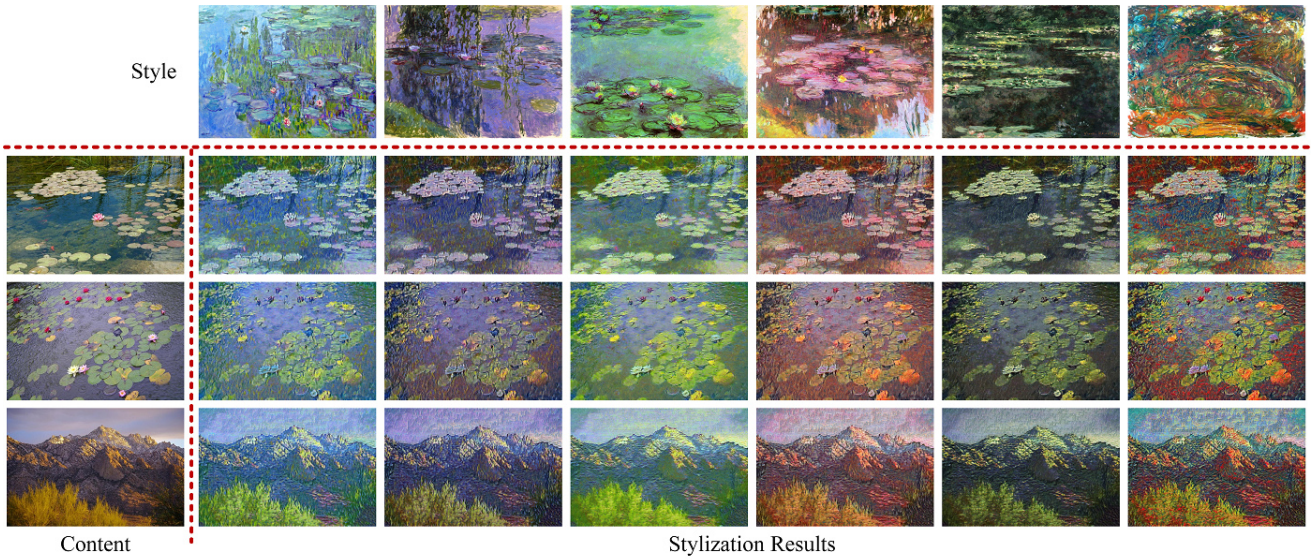

Figure 1: Stylization examples generated by our proposed DualAST. The first row shows several artworks from Claude Monet, which are taken as the style images. The first column shows the content images. The rest of the images are new artworks generated by our model.

Abstract

Artistic style transfer is an image editing task that aims at repainting everyday photographs with learned artistic styles. Existing methods learn styles from either a single style example or a collection of artworks. Accordingly, the stylization results are either inferior in visual quality or limited in style controllability. To tackle this problem, we propose a novel Dual Style-Learning Artistic Style Transfer (DualAST) framework to learn simultaneously both the holistic artist-style (from a collection of artworks) and the specific artwork-style (from a single style image): the artist-style sets the tone (i.e., *the overall feeling) for the stylized image, while the artwork-style determines the details of the stylized image, such as color and texture. Moreover, we introduce a Style-Control Block (SCB) to adjust the styles of generated images with a set of learnable style-control factors. We conduct extensive experiments to evaluate the performance of the proposed framework, the results of which confirm the superiority of our method.*

1. Introduction

Artistic style transfer aims to transfer the styles of artworks onto arbitrary photographs to create novel artistic images. The study on this topic has drawn much attention in recent years due to its scientific and artistic values. To achieve satisfying and convincing stylizations, it is very important to clarify the characteristics of artwork creation.

First, each artist has a specific painting style, which is distinguishable from others. We can get a good grasp of an

* Corresponding author

artist's style by studying the collection of his artworks. Intuitively, the more the number of artworks, the more profound and accurate our understanding of the artist's style. Second, there exist many variations within one style. For example, although *The Starry Night*, *Sunflowers*, and *Portrait of Dr. Gachet* are all van Gogh's artworks, they are significantly different in terms of color and texture.

To summarize, for the artworks of an artist, from the holistic perspective, they form the unique painting style of the artist; from the individual perspective, they are different from each other, exhibiting remarkable variations. Only if we take these two perspectives into consideration can we synthesize satisfying artistic images in style transfer. Unfortunately, existing style transfer methods only consider one of the two perspectives, which makes them unable to solve each other's problem. To be more specific, **on the one hand**, some methods [24, 15] only focus on the first perspective and learn the holistic artist-style from a collection of artworks, neglecting the variations among different images. The advantage of [24, 15] is that they can produce striking stylization results with pretty high quality. The defects of them are mainly in two aspects: 1) their stylization results are very limited in style controllability; 2) they can only produce one kind of stylizations, lacking variety. Kotovenko *et al*. [14] and Svoboda *et al*. [27] noticed the variations within one style and alleviated this issue by introducing a fixpoint triplet style loss and a two-stage peer-regularization layer, respectively. However, they still did not achieve satisfying controllable reference-guided stylizations, because they did not enforce an explicit style constraint between the reference (style) image and the stylized image. **On the other hand**, the other methods [8, 16, 11, 10, 20, 25, 30] only focus on the second perspective and learn the specific artwork-style from a single style example. The advantage is that they can produce controllable stylization results, while the defect is that they are significantly inferior to [24, 15, 14, 27] in visual quality, since they did not leverage the rich style information reserved in the artwork dataset. As [24] pointed out, it is insufficient to only use a single artwork, because it might not represent the full scope of an artistic style.

Motivated by the observations and analyses above, we propose a novel artistic style transfer framework, termed as DualAST, which learns simultaneously both the holistic artist-style (from a collection of artworks) and the specific artwork-style (from a single style image). Moreover, to better control the style of the generated image, we introduce a Style-Control Block (SCB) to our model. Although some existing arbitrary style transfer methods (*e.g*., AdaIN [10]) can also be used to perform controllable reference-guided image synthesis by exploring a statistical transformation from reference (style) image onto content image, we argue that it is inflexible and suboptimal to employ the manually defined statistics to control the style of the generated image. In comparison, our proposed SCB treats style information as a modulator rather than statistical data, and produces a set of learnable style-control factors to adjust the style of the generated image. In this way, our method is able to produce high-quality and style controllable stylization results with plenty of variations, as shown in Figure 1.

In summary, our main contributions are threefold:

- We propose a novel artistic style transfer framework DualAST that learns holistic artist-style and specific artwork-style simultaneously, yielding high-quality and style controllable stylization results with plenty of variations.
- We introduce a Style-Control Block (SCB) to better control the style of the generated image according to the input reference (style) image, resulting in artistically better stylizations.
- We demonstrate the effectiveness and strength of our approach by extensive comparisons with several state-of-the-art style transfer methods.

2. Related Work

Learning style from a single artwork. Recently, the seminal work of Gatys *et al*. [8] found that multi-level feature statistics extracted from a pre-trained Convolutional Neural Network (CNN) model can be used to separate content and style information, making it possible to recombine the content of a photograph with the style of a well-known artwork to create new artistic images. Since then, a series of follow-up works have been proposed to achieve better performance in many aspects, such as efficiency [11, 17, 28], quality [16, 31, 21, 37, 39, 35, 13], generalization [7, 4, 5, 10, 20, 25, 18, 22], and diversity [19, 29, 32, 33].

Efficiency. Although Gatys *et al*. [8] is able to generate visually stunning stylized images, its inference speed is prohibitively slow due to the iterative optimization process. To tackle this problem, [11, 17, 28] proposed to replace the optimization process with feed-forward neural networks, achieving real-time style transfer. ***Quality.*** Li *et al*. [16] improved the results of [8] by studying a combination of generative Markov random field (MRF) models and discriminatively trained deep convolutional neural networks (DCNNs). Ulyanov *et al*. [29] introduced an instance normalization module to replace batch normalization with significantly improvements to the quality of image stylization. Zhang *et al*. [39] introduced a more flexible and general universal style transfer technique that explicitly considers the matching of semantic patterns in content and style images. Yao *et al*. [37] developed an attention-aware multi-stroke style transfer model that is capable of generating striking

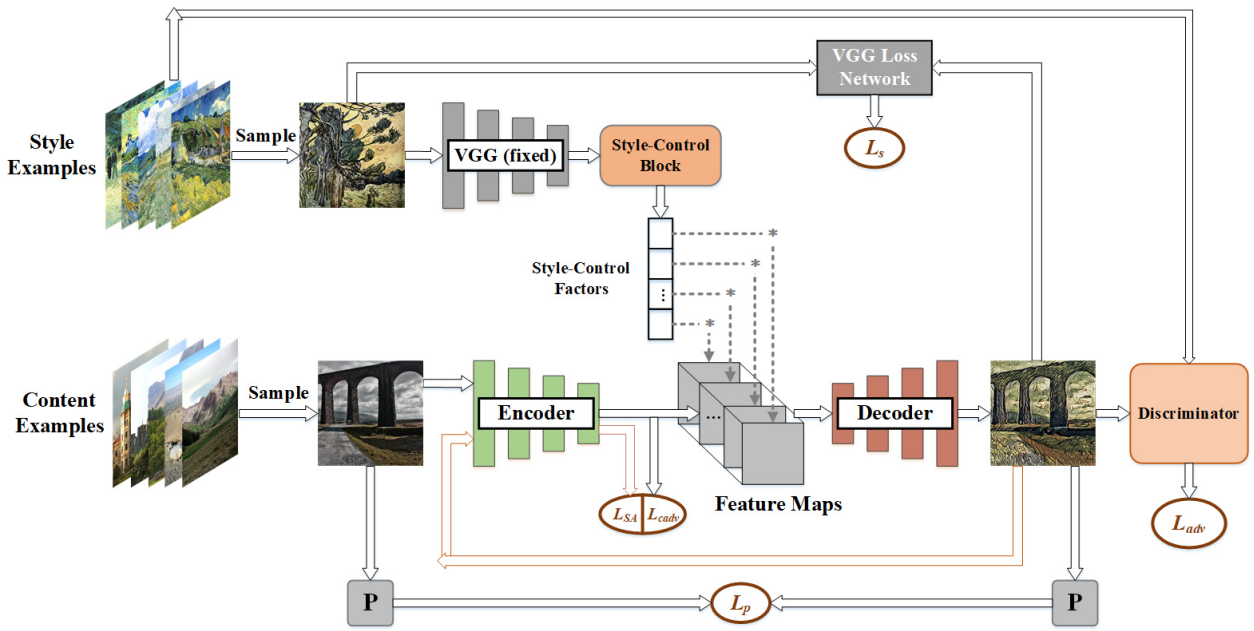

Figure 2: Illustration of the proposed DualAST framework. (1) The encoder captures style-aware content features from the content image with the constraint of the style-aware content loss $\mathcal{L}_{SA}$ and the style-aware content adversarial loss $\mathcal{L}_{cadv}$. (2) The style-control block takes the style features extracted by the fixed VGG network as its input, and produces a set of style-control factors to control the style of the generated image. (3) The decoder learns the holistic artist-style and the specific artwork-style with the constraint of the adversarial loss $\mathcal{L}_{adv}$ and the artwork-style loss $\mathcal{L}_s$. The soft reconstruction loss $\mathcal{L}_p$ is used to preserve the main content structure of the content image.

stylized images with multiple stroke patterns. ***Generalization.*** Universal style transfer methods have recently been proposed to transfer arbitrary styles through only one single model. Chen *et al*. [5] introduced a style swap operation to match and swap local patches between the intermediate features of content and style images, which for the first time realized universal style transfer. Huang *et al*. [10] proposed a novel adaptive instance normalization (AdaIN) layer that adjusts the mean and variance of the content input to match those of the style input. Li *et al*. [20] performed a pair of feature transforms, whitening and coloring (WCT), for feature embedding within a pre-trained encoder-decoder module. Sheng *et al*. [25] integrated a style decorator for semantic style feature propagation and an hourglass network for multi-scale holistic style adaptation. Li *et al*. [18] proposed a learnable linear transformation matrix which is conditioned on an arbitrary pair of content and style images. ***Diversity.*** Recently, diversified style transfer methods have been proposed to generate diverse stylization results, endowing users with more choices to select the satisfactory results according to their own preferences. Li *et al*. [19] and Ulyanov *et al*. [29] proposed to penalize the similarities of different stylization results in a mini-batch. Wang *et al*. [32] achieved better diversity by using an orthogonal noise matrix to perturb the image feature maps while keeping the original style information unchanged.

Despite the great progress, the image quality achieved by these methods is still insufficient, because they only focus on the limited style information contained in a single artwork.

Learning style from a collection of artworks. [24] argues that it is not enough to only use a single artwork, because it might not represent the full scope of an artistic style. Aiming at this problem, [24] proposed to learn style from a collection of artworks instead of a single style image, greatly improving the quality of stylized images. [15] introduced a novel content transformation block designed as a dedicated part of the network to alter an object in a content- and style-specific manner. [14] introduced two novel losses: a fixpoint triplet style loss to learn subtle variations within one style or between different styles and a disentanglement loss to ensure that the stylization is not conditioned on the real input photo. [27] provided a novel model to produce high-quality stylized images in the zero-shot setting and allow for more freedom in changes to the content geometry. Although [24, 15, 14, 27] achieved superior performance in terms of visual quality, they suffered from the uncontrollability of stylizations.

In this paper, we propose to learn simultaneously both the holistic artist-style from a collection of artworks and the

specific artwork-style from a single reference image, yielding high-quality and style controllable stylization results.

3. Proposed Method

Our goal is to learn an artistic style transfer model that captures both the holistic artist-style and the specific artwork-style. The holistic artist-style is learned from a collection of an artist's artworks, which sets the tone (*i.e.*, the overall feeling) for the stylized image. The specific artwork-style is learned from a single reference image, which determines the details (*i.e.*, color and texture) of the stylized image. To achieve this goal, we propose a novel DualAST framework, which we will introduce in details.

3.1. Dual Style Learning

Let X and Y be the sets of photos and artworks, respectively. As mentioned above, we aim to learn both the holistic artist-style from Y and the specific artwork-style from $y \in Y$, and then transfer them to an arbitrary content image $x \in X$ to create new artistic images. Figure 2 illustrates the pipeline of our proposed DualAST.

Holistic artist-style learning. We employ GAN (Generative Adversarial Network) [9, 23, 2, 38, 3] to align the distribution of generated images with artistic images in Y. The GAN consists of a generator $\mathcal{G}$ and a discriminator $\mathcal{D}_s$ that compete against each other. Note that for $\mathcal{G}$, we adopt an encoder-decoder architecture containing an encoder E and a decoder D. We express the adversarial loss as:

$$\begin{aligned}\mathcal{L}_{adv} := \mathop{\mathbb{E}}_{y\sim Y}[log(\mathcal{D}_s(y))]+ \\ \mathop{\mathbb{E}}_{x\sim X}[log(1-\mathcal{D}_s(D(E(x),\tau)))]\end{aligned} \tag{1}$$

where τ is the style-control factor used to control the style of the generated image (details in Section 3.2).

Specific artwork-style learning. Similar to previous works [8, 11, 17], we leverage a fixed pre-trained VGG-19 network [26] ϕ to compute the artwork-style loss. Inspired by the loss designations in [10, 22], we formulate the artwork-style loss $\mathcal{L}_s$ as,

$$\begin{aligned}\mathcal{L}_s := \sum_{i=1}^{n} \parallel \mu(\phi_i(D(E(x),\tau))) - \mu(\phi_i(y)) \parallel_2 + \\ \sum_{i=1}^{n} \parallel \sigma(\phi_i(D(E(x),\tau))) - \sigma(\phi_i(y)) \parallel_2\end{aligned} \tag{2}$$

where μ and σ are channel-wise mean and standard deviation, respectively. ϕ_i denotes a layer in VGG-19 used to compute the artwork-style loss. In our experiments we use relu1_1, relu2_1, relu3_1, relu4_1, and relu5_1 layers with equal weights.

Content structure preservation. Apart from learning the artistic style, style transfer also requires preserving the content structure of the content image x. To satisfy this requirement, we enforce a soft reconstruction loss between x and the stylization result $D(E(x),\tau)$,

$$\mathcal{L}_p := \mathop{\mathbb{E}}_{x\sim X}[\parallel P(D(E(x),\tau)) - P(x) \parallel_2^2] \tag{3}$$

where P is an average pooling layer. Compared with typical reconstruction loss, the soft reconstruction loss seeks to preserve only the essential content structure information rather than all the detailed information, which is more in line with the goal of style transfer (we demonstrate its effectiveness in the supplementary material via ablation studies).

Moreover, to empower E with the ability of capturing style-aware details from x, we adopt a style-aware content loss [24],

$$\mathcal{L}_{SA} := \mathop{\mathbb{E}}_{x\sim X}[\parallel E(D(E(x),\tau)) - E(x) \parallel_2^2] \tag{4}$$

However, a drawback of $\mathcal{L}_{SA}$ is that it is very sensitive to the value of its input. Thus, it can be easily minimized if E uses a small signal ($\parallel E(D(E(x),\tau)) \parallel \to 0, \parallel E(x) \parallel \to 0$), which reduces the loss but does not increase the similarity between $E(D(E(x),\tau))$ and $E(x)$. Aiming at this potential problem, we employ a feature discriminator $\mathcal{D}_f$ and introduce a style-aware content adversarial loss,

$$\begin{aligned}\mathcal{L}_{cadv} := \mathop{\mathbb{E}}_{x\sim X}[log(\mathcal{D}_f(E(x)))+ \\ log(1-\mathcal{D}_f(E(D(E(x),\tau))))]\end{aligned} \tag{5}$$

$\mathcal{L}_{cadv}$ seeks to minimize the distribution deviation, unrelated to the value of its input. This way, even if E uses a small signal, the similarity between $E(D(E(x),\tau))$ and $E(x)$ still can be promoted with $\mathcal{L}_{cadv}$.

Full objective. We summarize all aforementioned losses and obtain the full objective of our model,

$$\begin{aligned}\mathcal{L}_{full} := \lambda_{adv}\mathcal{L}_{adv} + \lambda_s\mathcal{L}_s + \lambda_p\mathcal{L}_p \\ +\lambda_{SA}\mathcal{L}_{SA} + \lambda_{cadv}\mathcal{L}_{cadv}\end{aligned} \tag{6}$$

where hyper-parameters λ_{adv}, λ_s, λ_p, λ_{SA}, and λ_{cadv} define the relative importance of the components in the overall loss function. They have been chosen by hand and will be shown below.

3.2. Style-Control Block

Existing arbitrary style transfer methods can be used to perform controllable reference-guided image synthesis by exploring a statistical transformation from reference image onto content image. For example, AdaIN [10] matches the

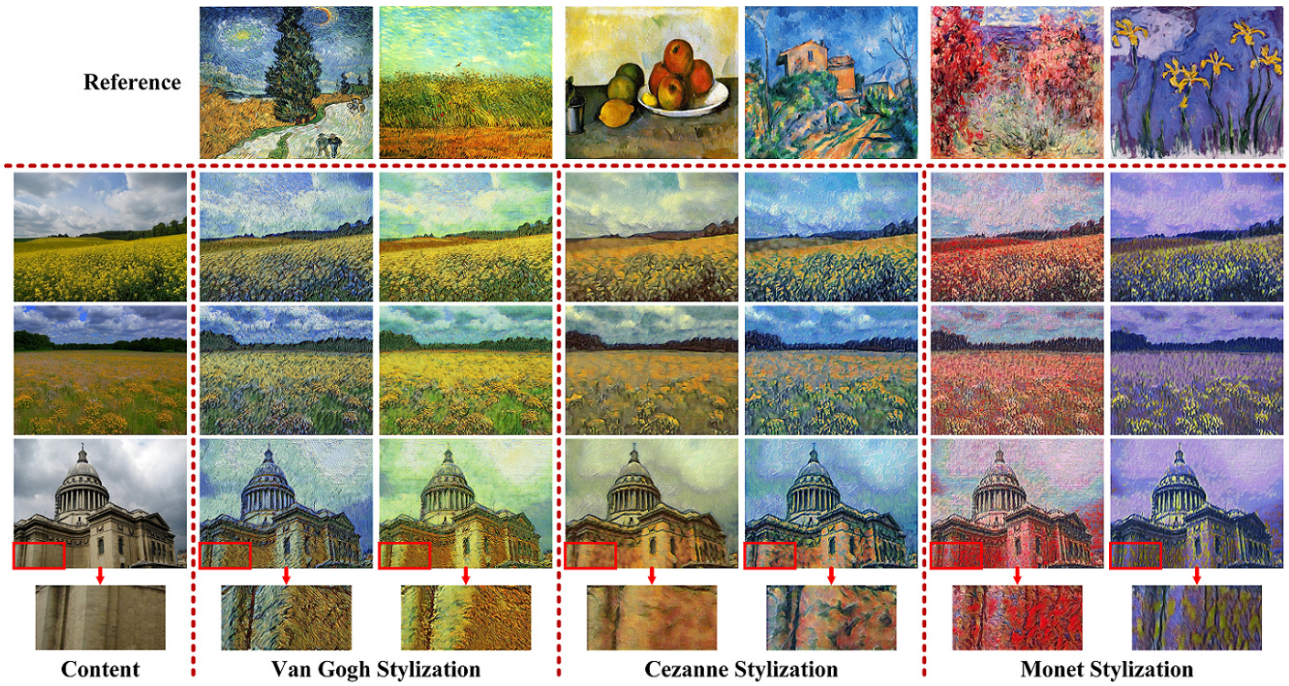

Figure 3: Stylization examples generated by our proposed DualAST. The first row shows several artworks from van Gogh, Cezanne, and Monet, which are taken as the style images. The first column shows the content images. The rest of the images are new artworks generated by our model.

means and variances of deep features between content and style images. However, we argue that it is inflexible and suboptimal to employ the manually defined statistics to control the style of the generated image.

In this work, we consider the reference-guided image synthesis problem from a different perspective and introduce a Style-Control Block (SCB) to our model. In detail, SCB takes the style features extracted from the reference image y as its input, and produces a set of learnable parameters τ (which we call style-control factors in this paper) containing the style characteristics of y,

$$\tau := SCB(\phi_{relu5\_1}(y)) \tag{7}$$

As shown in Figure 2, each scalar in $\tau \in \mathbb{R}^{B\times C}$ is used to control a feature map of $E(x) \in \mathbb{R}^{B\times H\times W\times C}$ via a simple scalar multiplication operation. In this way, the style characteristics of y can be injected to the decoder to guide the synthesis of the stylized image.

Compared with previous methods, our proposed SCB treats style information as a modulator rather than statistical data, exhibiting two advantages: i) complicated statistics computation is not required; ii) the style-control factors in SCB can be adaptively changed in a learnable manner according to different style examples, more flexible and reasonable than manually defined statistics.

3.3. Implementation Details

We adapt the architectures for our encoder and decoder from [11] which has shown impressive results for artistic style transfer. Specifically, the encoder E is composed of 1 stride-1 and 4 stride-2 convolution layers. The decoder D contains 9 residual blocks, 4 upsampling blocks, and 1 stride-1 convolution layer. The style-control block SCB has 2 stride-2 and 1 stride-1 convolution layers. The style discriminator $\mathcal{D}_s$ is a fully convolutional networks with 7 stride-2 convolution layers. The feature discriminator $\mathcal{D}_f$ includes 3 stride-2 convolution layers and 1 fully connected layer. As for P, it is an average pooling layer. The loss weights in Equation (6) are set to $\lambda_{adv} = 1$, $\lambda_s = 0.3$, $\lambda_p = 100$, $\lambda_{SA} = 100$, $\lambda_{cadv} = 10$. The learning rate is initially set to 0.0002 and then decreased by a factor of 10 after 200000 iterations (300000 iterations in total). Our code is available at: https://github.com/HalbertCH/DualAST.

4. Experimental Results

To evaluate the proposed method, extensive experiments and comparisons have been conducted. First, in Section 4.1, we show the high-quality and style controllable stylization results generated by our model and perform qualitative comparisons. Next, quantitative results are presented

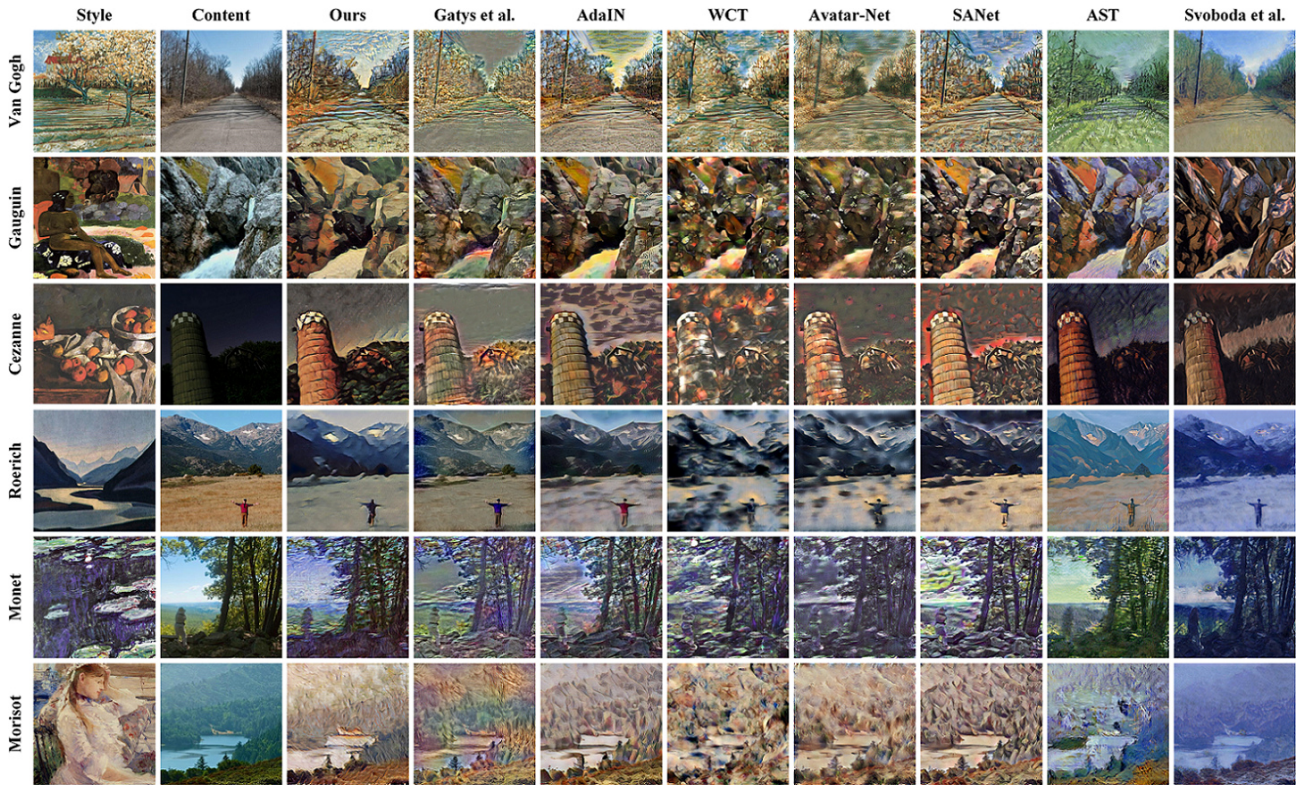

Figure 4: Qualitative comparisons. The first column shows the style images from different artists. The second column shows the content images. The rest columns show the stylization results generated by different style transfer methods.

in Section 4.2. Finally, in Section 4.3, we conduct comprehensive ablation studies to validate the effectiveness of each component in our model.

Baselines. We use Gatys *et al*. [8], AdaIN [10], WCT [20], Avatar-Net [25], SANet [22], AST [24], and Svoboda *et al*. [27] as our baselines. Among them, Gatys *et al*. [8], AdaIN [10], WCT [20], Avatar-Net [25], and SANet [22] learn style from a single style image, while AST [24] and Svoboda *et al*. [27] learn style from a collection of artworks. All the baselines are trained using publicly available implementations with default configurations.

Datasets. Following [24, 15, 14, 27], we use Place365 [41] and WikiArt [12] for content and style images, respectively. During training, we randomly sample image patches of size 768×768 from Place365 and WikiArt. Note that in the inference stage, both the content and style images can be of any size.

4.1. Qualitative Results

In Figure 3, we show our stylization results based on different artworks of different artists, including Vincent van Gogh, Paul Cezanne, and Claude Monet. It can be seen that the stylized images exhibit both remarkable visual quality and satisfying style controllability. From the zoom-in part in Figure 3, we further observe that for the stylized images based on the same artist's style, despite different colors and style patterns, their strokes are very similar; while for the stylized images based on different artists' styles, their strokes are clearly distinguishable from each other. For example, the van Gogh stylizations all employ curved and interlaced strokes, while the Cezanne stylizations all employ large flat strokes.

To validate the superiority of our method, we compare our stylization results with those of aforementioned 7 baselines in Figure 4. Gatys *et al*. [8] is likely to encounter a bad local minimum (*e.g*., rows 1, 5, and 6). AdaIN [10] usually introduces undesired colors and patterns that do not exist in style images (*e.g*., rows 1, 2, and 5). WCT [20] fails to preserve the main content structures, resulting in messy and less-structured stylizations (*e.g*., rows 3, 4, and 6). Avatar-Net [25] sometimes blurs the semantic structures (*e.g*., rows 1, 4, and 6). SANet [22] tends to apply repeating style patterns to stylized images (*e.g*., rows 2, 3, and 6). Above five style transfer methods all learn style from a single style image, resulting in insufficient visual quality. As for AST [24] and Svoboda *et al*. [27], they are able to produce stunning stylized images. However, as we can see from the last two columns, their stylization results have very limited rel-

Table 1: The deception rate and user study (in terms of visual quality and style controllability) for different methods. The higher the better. The best scores are reported in bold.

	Gatys *et al.*	AdaIN	WCT	Avatar-Net	SANet	AST	Svoboda *et al.*	DualAST
Deception Rate	0.206	0.065	0.027	0.046	0.122	0.454	0.278	**0.589**
Visual Quality (%)	0.061	0.052	0.011	0.034	0.093	0.253	0.191	**0.305**
Style Controllability (%)	**0.194**	0.117	0.154	0.121	0.168	0.005	0.059	0.182

evance to style images. This is because they only learn the holistic artist-style from the whole artwork dataset, resulting in uncontrollable stylizations.

In comparison, our proposed DualAST learns simultaneously both the holistic artist-style (from the artwork dataset) and the specific artwork-style (from a single style image). The results in the third column of Figure 4 demonstrate the effectiveness and superiority of our method. More results can be found in our supplementary material.

4.2. Quantitative Results

Assessing artistic style transfer results could be a highly subjective task. In this section, we adopt two quantitative evaluation metrics: deception rate [24] and user study, to better evaluate our method.

Deception rate. This metric was introduced by Sanakoyeu *et al.* [24] to quantitatively and automatically assess the quality of stylization results. First, [24] trained a VGG-16 network to classify 624 artists on WikiArt [12] from scratch. Then, the pre-trained network was employed to predict which artist the stylized image belongs to. Finally, the deception rate was calculated as the fraction of times that the network predicted the correct artist. We report the deception rate for the proposed DualAST and seven baseline models in the second row of Table 1. It can be observed that our method achieves the highest score and outperforms the other methods by a large margin.

User study. User study has been widely adopted by previous works [22, 24, 34, 6, 1, 36, 40] to investigate user preference over different visual results. Here we conduct two user studies to evaluate the user preference of our and competing methods in terms of visual quality and style controllability, respectively.

Visual quality. Given various photographs, we stylize them in the style of different artists using DualAST and 7 baseline methods. Then we show the randomly ordered stylized images produced by 8 compared methods to participants and ask them to select the image that best represents the style of the target artist. We finally collect 1000 votes from 50 participants. We report the percentage of votes for each method in the third row of Table 1, where we can see that the stylization results obtained by our method are preferred more often than those of other methods.

Style controllability. We select 20 content and style image pairs and take them as the inputs of above 8 compared methods, yielding 20 stylized images for each method. Then we show these stylized images alongside the style image and content image to participants and ask them to choose the image that learns the most characteristics from the style image. We report the percentage of votes for each method in the fourth row of Table 1. We observe that DualAST achieves the second-highest score, after Gatys *et al.* [8] The reason behind this is that Gatys *et al.* [8] is an optimization-based method that performs an optimization process for every style, while the other methods train one single model to transfer arbitrary styles.

To summarize, existing artistic style transfer methods are either inferior in visual quality or limited in style controllability. In comparison, our DualAST achieves both remarkable visual quality and satisfying style controllability.

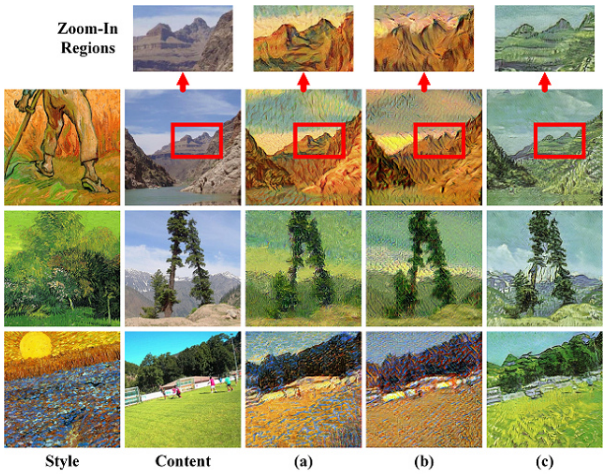

Figure 5: Ablation results for holistic artist-style and specific artwork-style learning. (a) The results of full DualAST. (b) The results of DualAST w/o holistic artist-style learning. (c) The results of DualAST w/o specific artwork-style learning.

4.3. Ablation Studies

In this section, we explore each component's effect in DualAST and validate their importance by ablation studies.

With and without holistic artist-style learning. Here we train a DualAST model that does not involve holistic

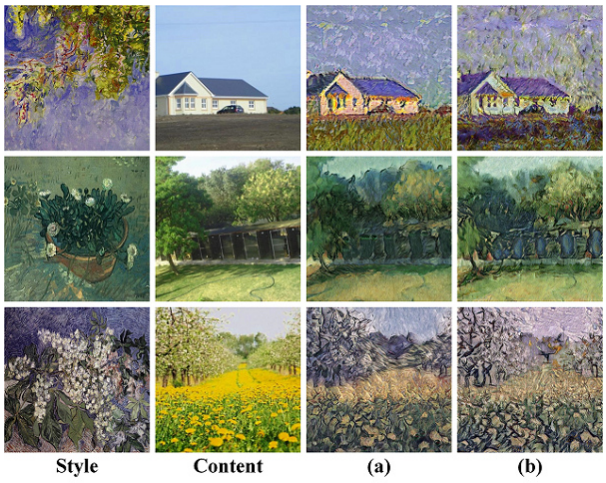

Figure 6: Ablation results for the style-aware content adversarial loss. (a) The results of full DualAST. (b) The results of DualAST w/o the style-aware content adversarial loss.

artist-style learning. The experimental results are shown in Figure 5. We can see that, compared with the full DualAST model, the DualAST model without holistic artist-style learning produces less appealing stylization results with some noticeable artifacts (see the zoom-in regions in row 1, the strokes of (a, c) are much more natural and smooth than the stroke of (b)). The reason could be that it is insufficient to learn style from a single artwork, because it might not represent the full scope of an artistic style. To enhance the visual quality of the stylized image with more holistic artist-style (for example, stroke), it is important to leverage the rich style information reserved in the whole artwork dataset.

With and without specific artwork-style learning. Similarly, we also train a DualAST model that does not involve specific artwork-style learning. As shown in Figure 5 (c), its stylization results have little relevance to the style (reference) image. The reason behind this is that the model only focuses on learning the holistic artist-style from the whole artwork dataset, neglecting the variations among different artworks. To achieve controllable reference-guided stylizations, it is important to learn more specific artwork-style (*e.g.*, color and texture) from the reference image.

With and without the style-aware content adversarial loss. As discussed in Section 3.1, we introduced a style-aware content adversarial loss $\mathcal{L}_{cadv}$ to alleviate the limitation of the style-aware content loss [24] $\mathcal{L}_{SA}$. In Figure 6, we show stylizations of our method with and without $\mathcal{L}_{cadv}$. We find that our full model better matches the target style to the content image, yielding visually more pleasing results. Take the first row as an example, there are some content distortions in the ground of image (b), while image (a) does not have such problems.

With and without the style-control block. To investigate the effect of our proposed style-control block (SCB), we substitute it with a concatenation layer (which directly concatenates the content and style features) and an AdaIN [10] layer, respectively. The experimental results are shown in Figure 7. From the zoom-in parts we can see that the concatenation layer tends to preserve the content structures of the style image, while the AdaIN layer fails to learn the target color distribution and introduces some unwanted artifacts. Such problems can be avoided with our SCB.

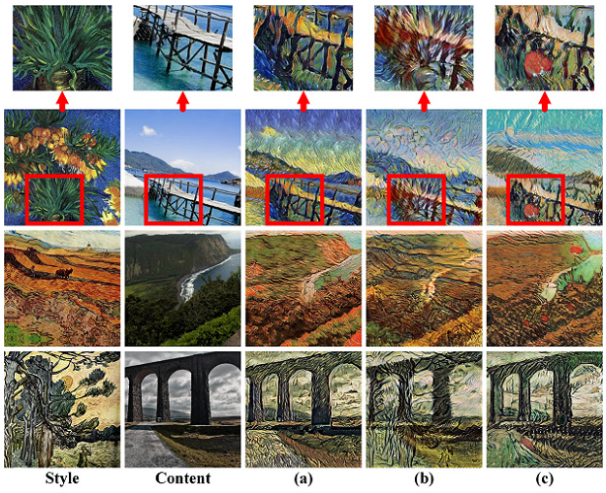

Figure 7: Ablation results obtained by applying (a) the style-control block, (b) the concatenation layer, and (c) the AdaIN layer.

5. Conclusion

In this paper, we propose a novel style transfer framework, termed as DualAST, to address the artistic style transfer problem from a new perspective. The core idea of DualAST is to learn simultaneously both the holistic artist-style and the specific artwork-style: the first style sets the tone for the stylized image, while the second style determines the details of the stylized image. Furthermore, we introduce a Style-Control Block (SCB) to adjust the styles of generated images with a set of learnable style-control factors. Extensive experimental results demonstrate the strength of our approach against the state-of-the-art in terms of visual quality and style controllability. As a future direction, we will further explore the more detailed relation between the holistic artist-style and the specific artwork-style for enhanced stylizations.

Acknowledgments. This work was supported in part by National Key R & D Plan Project (No: 2020YFC1523201, 2020YFC1523101, 2020YFC1522701), Zhejiang Science and Technology Project (No: 2019C03137), and Zhejiang Fund Project (No: LY19F020049).

References

[1] Ivan Anokhin, Pavel Solovev, Denis Korzhenkov, Alexey Kharlamov, Taras Khakhulin, Aleksei Silvestrov, Sergey Nikolenko, Victor Lempitsky, and Gleb Sterkin. High-resolution daytime translation without domain labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7488–7497, 2020. 7

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 4

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 4

[4] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1897–1906, 2017. 2

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

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 7

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

[8] 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, 4, 6, 7

[9] 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. 4

[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, 4, 6, 8

[11] 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, 4, 5

[12] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. *arXiv preprint arXiv:1311.3715*, 2013. 6, 7

[13] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019. 2

[14] 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. 2, 3, 6

[15] 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. 2, 3, 6

[16] 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

[17] 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, 4

[18] 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, 3

[19] 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

[20] 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, 6

[21] Ming Lu, Hao Zhao, Anbang Yao, Yurong Chen, Feng Xu, and Li Zhang. A closed-form solution to universal style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5952–5961, 2019. 2

[22] 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, 4, 6, 7

[23] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 4

[24] 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. 2, 3, 4, 6, 7, 8

[25] 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, 6

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[27] 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. 2, 3, 6

[28] 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

[29] 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

[30] Huan Wang, Yijun Li, Yuehai Wang, Haoji Hu, and Ming-Hsuan Yang. Collaborative distillation for ultra-resolution universal style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1860–1869, 2020. 2

[31] 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. 2

[32] 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

[33] Zhizhong Wang, Lei Zhao, Haibo Chen, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Diversified patch-based style transfer with shifted style normalization. *arXiv preprint arXiv:2101.06381*, 2021. 2

[34] 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. 7

[35] 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

[36] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. Stylization-based architecture for fast deep exemplar colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9363–9372, 2020. 7

[37] 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

[38] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018. 4

[39] 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

[40] 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

[41] 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. 6