

Distribution Aligned Multimodal and Multi-domain Image Stylization

MINXUAN LIN, NLPR, Institute of Automation, Chinese Academy of Sciences & School of Artificial Intelligence, University of Chinese Academy of Sciences

FAN TANG, School of Artificial Intelligence, Jilin University

WEIMING DONG, NLPR, Institute of Automation, Chinese Academy of Sciences & CASIA-LLvision Joint Lab

XIAO LI, Microsoft Research Asia

CHANGSHENG XU, NLPR, Institute of Automation, Chinese Academy of Sciences & CASIA-LLvision Joint Lab

CHONGYANG MA, Kuaishou Technology

Multimodal and multi-domain stylization are two important problems in the field of image style transfer. Currently, there are few methods that can perform multimodal and multi-domain stylization simultaneously. In this study, we propose a unified framework for multimodal and multi-domain style transfer with the support of both exemplar-based reference and randomly sampled guidance. The key component of our method is a novel style distribution alignment module that eliminates the explicit distribution gaps between various style domains and reduces the risk of mode collapse. The multimodal diversity is ensured by either guidance from multiple images or random style codes, while the multi-domain controllability is directly achieved by using a domain label. We validate our proposed framework on painting style transfer with various artistic styles and genres. Qualitative and quantitative comparisons with state-of-the-art methods demonstrate that our method can generate high-quality results of multi-domain styles and multimodal instances from reference style guidance or a random sampled style.

CCS Concepts: • **Applied computing** → **Fine arts**; • **Computing methodologies** → *Appearance and texture representations*

Additional Key Words and Phrases: Image stylization, multi-domain, multimodal

This work was supported by National Key R&D Program of China under Grant No. 2020AAA0106200, and by National Natural Science Foundation of China under Grants No. 61832016 and No. U20B2070.

Authors' addresses: M. Lin, NLPR, Institute of Automation, Chinese Academy of Sciences and School of Artificial Intelligence, University of Chinese Academy of Sciences, #95 East Zhongguancun Road, Haidian District, Beijing, 100190, China; email: linminxuan2018@ia.ac.cn; F. Tang, School of Artificial Intelligence, Jilin University, #2699 Qianjin Street, Changchun, Jilin, 130012, China; email: tangfan@jlu.edu.cn; W. Dong and C. Xu, NLPR, Institute of Automation, Chinese Academy of Sciences and CASIA-LLvision Joint Lab, #95 East Zhongguancun Road, Haidian District, Beijing, 100190, China; emails: weiming.dong@ia.ac.cn, csxu@nlpr.ia.ac.cn; X. Li, Microsoft Research Asia, #5 Dan Ling Street, Haidian District, Beijing, 100080, China; email: pableetoli@gmail.com; C. Ma, Kuaishou Technology, #6 Shangdi West Road, Haidian District, Beijing, 100080, China; email: chongyangm@gmail.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

1551-6857/2021/07-ART96 \$15.00

<https://doi.org/10.1145/3450525>

ACM Reference format:

Minxuan Lin, Fan Tang, Weiming Dong, Xiao Li, Changsheng Xu, and Chongyang Ma. 2021. Distribution Aligned Multimodal and Multi-domain Image Stylization. *ACM Trans. Multimedia Comput. Commun. Appl.* 17, 3, Article 96 (July 2021), 17 pages.
<https://doi.org/10.1145/3450525>

1 INTRODUCTION

Style transfer is a typical technique to stylize content images in the style of another input [7, 9–11]. Recently, **image-to-image (I2I)** translation methods based on conditional generative adversarial networks [14] have played a pivotal role in addressing the problem of style transfer. While these pioneering techniques have shown promising results for generating a single stylized output from a reference image, two interesting problems, namely, *multi-domain* and *multimodal* stylization, have been raised. Multi-domain stylization methods seek for improved *controllability* during the style transfer process, i.e., to generate different styles based on guidance from multiple domains. By contrast, multimodal methods focus on the *diversity* of generated stylization results, i.e., to synthesize multiple different results that are all consistent with the same reference style.

Majority of existing multi-domain methods, e.g., StarGAN [5], are inherited from unpaired conditional I2I translation [45], which learns one-to-one mapping between two domains, and thus loses the ability of synthesizing multimodal results. Multimodal methods, such as MUNIT [13], are usually limited to handling only two domains at one time and do not support multi-domain stylization. Although several methods [34, 42] have been proposed to address multimodal and multi-domain stylization simultaneously, they are restricted to generating images with only random sampling (i.e. generating results by random sampling from the style space) or suffer the issue of mode collapse due to the use of Kullback-Leibler divergence.

The key to addressing these problems is to construct a style embedding space that (1) preserves style information from reference style images for exemplar-based multimodal stylization; (2) is smooth enough for random sampling-based multimodal stylization; (3) has a uniformly covered style distribution for avoiding mode collapse; and (4) provides flexible control using domain labels for multi-domain stylization.

To overcome these challenges, we propose a unified framework that achieves multi-domain and multimodal stylization simultaneously with exemplar-based guidance and random sample guidance and reduces the possibility of mode collapse. The aligned space has the following properties: (1) a reference style can be extracted from a trained encoder to support exemplar-based stylization; (2) each conditioned space can support multimodal stylization via random sampling; (3) style features are sufficient to cover the sampling space; and (4) the space is conditioned on domains so that multi-domain stylization is also available. We demonstrate the strength of our method on painting style transfer with various artistic styles and genres. Qualitative and quantitative comparisons with state-of-the-art methods indicate that our approach can generate high-quality results of multi-domain and multimodal stylization.

2 RELATED WORK

Style transfer. Gatys et al. [11] first adopted convolution neural networks to deal with a single image stylization problem by using an iterative optimization procedure. For additional diversity, several arbitrary style transfer methods were proposed. WCT [22] progressively repeated whitening and coloring operations at multiple scales to alter any style patterns. Huang et al. [12] used the **adaptive instance normalization (AdaIN)** layer to align with the feature statistics of content and style images. AvatarNet [36] proposed a style decorator to make up the content feature

semantically with the style feature in multiscale layers. Sanakoyeu et al. [35] emphasized the style-aware content constraint to achieve real-time HD style transfer. Kotovenko et al. [18] used a content transformation module to focus on details. Li et al. [21] introduced a transformation matrix to transfer style across different levels flexibly and efficiently. Kotovenko et al. [17] exchanged the content and style of stylized images to disentangle the two elements for an improved style mix. Moreover, several attention-aware fashions [8, 33, 41], where the models learnt to adjust the influencing factor of the style feature for the content feature, were proposed. However, none of these above methods, which lack the controllability, support multi-domain style transfer.

I2I translation. Closely related to style transfer, I2I translation addresses a more general synthesis problem that shifts the style distribution from one domain to another while maintaining semantic features between images [14, 31, 44]. CGAN [31] rendered primitive translation processes with a noise condition. Pix2Pix [14] used conditional generative adversarial networks to transfer images between two domains. These methods were further improved by CycleGAN [45], which used a dual-learning approach and eliminated the requirement of paired data. While showing promising results, these methods were intrinsically limited to learn a mapping between two domains.

Based on these explorations, several methods attempted to address either multi-domain or multimodal I2I translation. ACGAN [32] proposed to append an auxiliary classifier in the discriminator to support multi-domain generation. For multi-domain translation, ComboGAN [1] leveraged multiple encoder-decoders for altering between different styles. StarGAN [5] used a unified conditional generator for multi-domain synthesis. SGN [4] explored the influence of mixed domains. For multimodal generation, MSGAN [29] introduced a new constraint that emphasized on the ratio of the distances between images and their corresponding latent codes. EGSC-IT [27] controlled the AdaIN parameters of image generator by using a style coding branch. FUNIT [25] dealt with multi-domain translation in a few-shot setting. However, not all of them can perform multimodal and multi-domain translation simultaneously.

Recently, several methods [6, 20, 26, 34, 40, 42] were proposed to achieve multi-domain and multimodal synthesis within a single framework. SMIT [34] used a combination of random noise and domain condition as guidance for image translation. DMIT [42] separated content, style, and domain information with different encoders. However, the limitations of guidance way (only support random sampling) and the difficulty of controlling style space by KL-divergence have become their obstacles. Concurrent to our work, StarGAN v2 [6] used a mapping network to transform a latent code to a style code from multiple domains. The multibranch strategy was also adopted by discriminators. As a result, the number of parameters inevitably increases with the addition of domains.

Disentangled latent representations. Building the mapping between the latent space and the image space promotes the quality and controllability of synthesized output. VAE [16] used reparametrization tricks to construct the relationships between the two spaces. CVAE [37] took one-hot label as the condition to construct multiple clusters in the latent space. AAE [28] proposed an adversarial strategy to force the latent space distribution to be close to the prior distribution. UNIT [24] adopted double VAEs to encode the latent vectors into a shared latent space. MUNIT [13] and DRIT [19] further disentangled the content feature and style feature into disparate manifolds. To disentangle multi-domain features, UFDN [23] aligned domain representations by an adversarial domain classifier. Kotovenko et al. [17] used fixpoint loss to decouple the content and style space. Similarly, we also encode the two properties by respective encoders.

Table 1 summarizes the different properties of our method and other related techniques. Most existing methods focus on either multi-domain or multimodal synthesis. Few of them

Table 1. Comparisons with Recent Methods for Image-to-image Translation and Style Transfer

Method	Unified Generator	Multiple modals	Multiple domains	Sample Guided	Exemplar Guided	Feature Adaptation
Pix2Pix	✓	—	—	—	—	—
CycleGAN	—	—	—	—	—	—
UNIT	—	—	—	—	—	—
StarGAN	✓	—	✓	—	—	—
MUNIT	—	✓	—	✓	✓	—
DRIT	—	✓	—	✓	✓	✓
UFDN	✓	—	✓	✓	—	✓
EGSC-IT	✓	✓	—	—	✓	—
DMIT	✓	✓	✓	✓	✓	—
SMIT	✓	✓	✓	✓	—	—
Ours	✓	✓	✓	✓	✓	✓

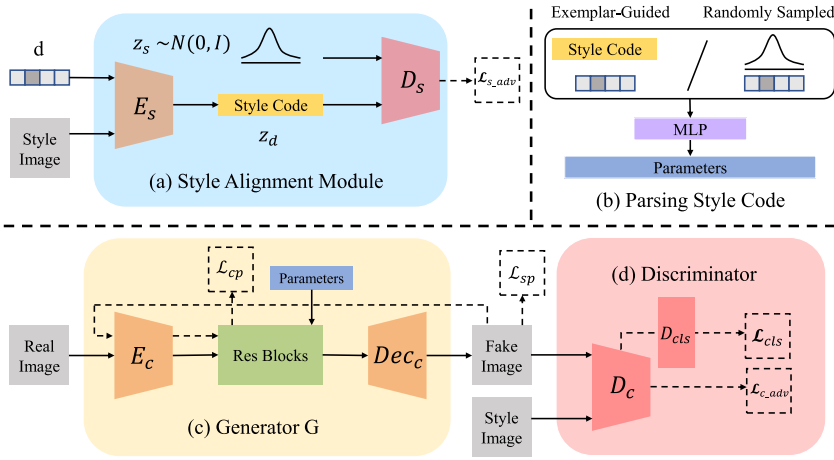


Fig. 1. Illustration of our framework.

have explored both with limited support of style guidance. However, the demand for model controllability and diversity is increasing. Various guidance approaches, such as exemplar guidance or random sampling guidance, bring rich maneuverability. In addition, each pair of cross-domain translation consumes considerable time and space to train an independent model. Thus, we propose a convenient fashion to achieve multimodal and multidomain stylization simultaneously.

3 OUR METHOD

The input of our method includes a natural *content image* $x \in \mathcal{X}$ that user wants to stylize, as well as a *style code* z associated with its domain label $d \in \mathcal{D}$, i.e., a one-hot vector indicating its style domain. The style code can either be generated from a reference *style image* $y_d \in \mathcal{Y}_d$ of a certain style domain label d , or directly sampled from a normal Gaussian distribution $\mathcal{N}(0, I)$. In case a style image is provided, the corresponding style code is extracted from our style encoder (Section 3.1). The style code z and the style domain label d are then converted to parameters that control the AdaIN layers of our image generator (Section 3.1). The output image $\tilde{y}_d$ is finally synthesized by our image generator on the basis of the content image x (Section 3.2) and the above style information. Figure 1 illustrates our entire framework.

3.1 Style Space Embedding

The key to integrating multimodal and multi-domain style transfer into a unified framework, without losing ability for either exemplar-guided or random sampling, is an embedded style space

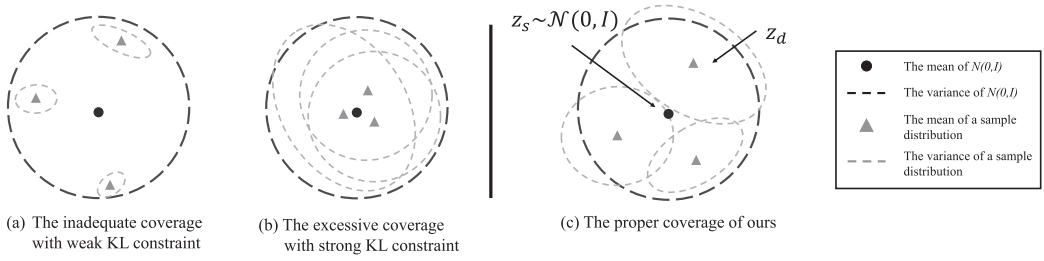


Fig. 2. Comparison of KL loss and our style alignment module. The improper KL constraint will lead to excessive or inadequate coverage. We avoid this situation by adversarial training to achieve complete and disjoint coverage.

that can be controlled at the inter-domain level and randomly traversed at the intra-domain level. In other words, the style space should (1) be clearly separated between different styles via domain label control and (2) form a smooth space that can be interpolated within a given style domain. To this end, we design a *style alignment module* for style space embedding. The embedded style code is then further converted into parameters to control the AdaIN layer of our image generation network as in Reference [12].

Style alignment module. Our style alignment module is an encoder-decoder network that constructs the embedded style space from style images of multiple domains. As shown in Figure 1(a), we feed a style image y_d and its corresponding domain label d into the style encoder E_s together to form a one-dimensional style embedding z_d . The style encoder E_s consists of multiple down-sampling blocks, and global average pooling is applied to the final layer to squeeze output style features.

In contrast to CVAE-GAN [3], our style alignment module does not have to reconstruct a given style image accurately. Instead, the goal of our style alignment module is to eliminate the explicit distribution gaps among various style domains and align them, i.e., different artist styles are controlled by d , and style space w.r.t each domain is aligned to Gaussian distribution to enable sampling and smooth interpolation. Thus, we avoid the reconstruction loss in Reference [3], because pixel-level reconstruction loss causes the encoding of extra content-related information into the style code, thus obstructing our goal of extracting only style information.

Additionally, the KL-divergence loss used in CVAE-GAN without reconstruction constraint leads to a trivial solution. Figures 2(a) and 2(b) show two situations when KL-divergence loss is under an inappropriate weight in one artist domain: (1) A weak KL constraint makes the variance of style feature promptly converge to zero and result in inadequate coverage. (2) A strong KL constraint makes the style feature to be indistinguishable from $N(0, I)$ and results in excessive coverage. That is, the style space is destroyed in both situations. To validate our analysis, we adopt KL-divergence to construct the style space constrained by different weights (from $1e-2$ to $1e-8$) in the same framework. We estimate the **inception score (IS)** and **learned perceptual image patch similarity (LPIPS)** score separately in our fixed test environment, and the results are shown in Figures 3(a) and 3(b). We plot the mean and variance under several weight configurations, which are calculated by different artist domains. The dash line represents the average performance in each metric. From Figure 3, we could conclude the following: (1) We have a high probability to improve performance when the weight close to the middle area ($1e-5$, $1e-6$) is chosen. By contrast, the performance worsens, because we adopt too large or too small hyper-parameters. (2) The performance of constructing style space through KL divergence has great fluctuation.

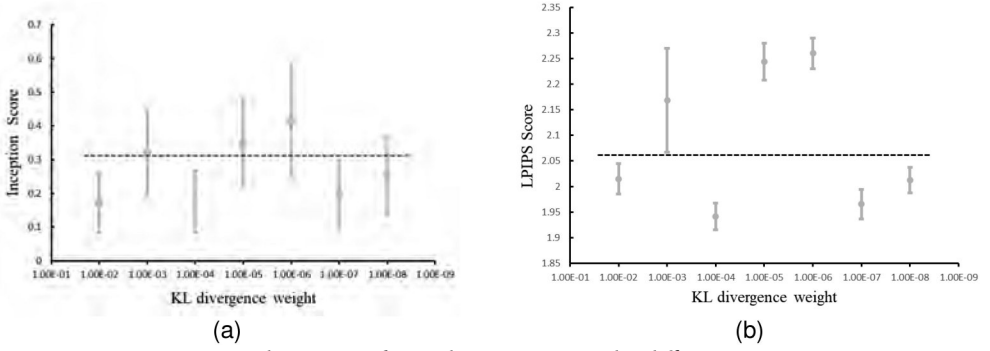


Fig. 3. Mean and variance of IS and LPIPS score under different KL constraints.

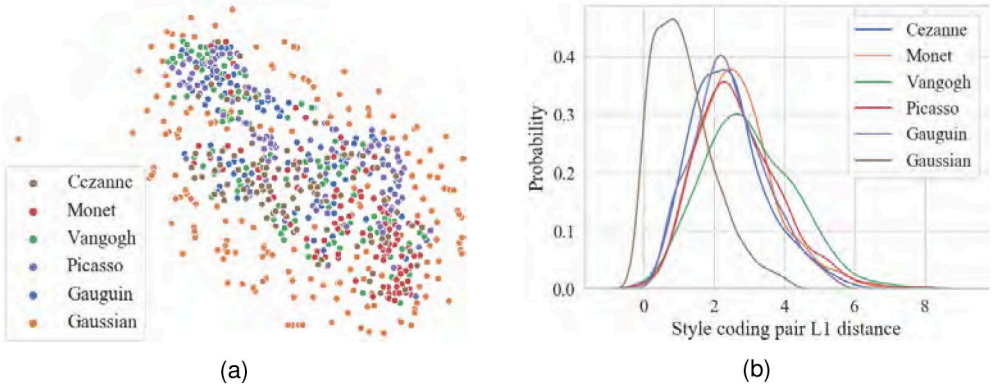


Fig. 4. (a) The t-SNE embedding visualization of style features from different artist domains. (b) L1 distance distribution of style coding pairs. To show the effect of alignment, two style features from the same domain are extracted by E_s as a pair to calculate the distribution of Manhattan distance.

To avoid unstable style space distribution, we remove the KL-divergence term and train our style alignment module with only *style adversarial loss* $\mathcal{L}_{s_adv}$:

$$\begin{aligned} \mathcal{L}_{s_adv} = & \mathbb{E}_z[\log D_s(z_s)] \\ & + \mathbb{E}_{y_d, d}[\log(1 - D_s(E_s(y_d, d)))], \end{aligned} \quad (1)$$

where z_s is randomly sampled from a Gaussian distribution $\mathcal{N}(0, I)$. The style alignment discriminator D_s determines whether the unknown style feature points z are from a Gaussian distribution or generated by E_s .

The adversarial loss tends to align the joint distribution of all domain styles (i.e., the unconditioned style space) to a Gaussian distribution. Consequently, each feature point from the same artist domain is arranged accordingly to cover different regions, and their union spans the full Gaussian distribution, as illustrated in Figure 2(c). A real case of the aligned feature distribution of our trained style space is illustrated in Figure 4(a) via t-SNE visualization and in Figure 4(b) via the distance distribution of the L1-distance between random sampled style feature pairs. Apparently, our style space has a complete coverage that consists of disjoint feature points.

Controllable image synthesis. Our style alignment module provides two possible ways to parse a style code, i.e., *exemplar guided* and *randomly sampled*. As shown in Figure 1(b), exemplar-guided style code is extracted from the style alignment encoder, providing precise control information.

Moreover, the corresponding stylized result is expected to achieve the same color distribution and texture appearance as the style image y_d . Randomly sampled style codes enable multimodal stylization in a certain domain by sampling the style code z_s and an arbitrary domain label d . Similar to Reference [13], we use the style code z and its corresponding domain label d for stylized image generation by controlling the parameters of the AdaIN layers. The style code z is concatenated with the style label d and transformed into channel-wise feature scale and bias for the AdaIN layer by using a multilayer perception network.

3.2 Stylized Image Generation

Image generator. Given a content image x and controlling parameters from z and d , the output image $\tilde{y}_d$ is synthesized by the stylized image generator G . Figure 1(c) illustrates our generator framework. Inspired by CycleGAN [45], our network is constructed by an encoder-decoder architecture that contains several down-sampling layers, residual blocks and up-sampling layers. In contrast to other I2I translation and stylization methods [13, 19], which use consistent normalization methods for most layers, we use Instance Normalization (IN) [39] in the down-sampling layers and the first half of residual blocks, AdaIN in the second half of residual blocks, and Layer Normalization (LN) [2] in the up-sampling layers to avoid irregular artifacts.

Loss functions. The goal of our stylized image generator network is to generate images that preserve fidelity with the original content image x and consistency with the style code z . To ensure that the stylized output image preserves the semantic content of the content image x , we use a *content preserving loss*, which constrains the output stylized image to achieve the same encoded content feature as input content image:

$$\mathcal{L}_{cp} = \|E_c(G(x, z, d)) - E_c(x)\|_1, \quad (2)$$

where the output $G(x, z, d)$ is a stylized image $\tilde{y}_d$. L1 distance is used as the metric.

To ensure the consistency between style image y_d and the synthesis output $\tilde{y}_d$ during the training process, we apply a *style preserving loss*, which computes the L1 distance of the gram matrix on the multiscale feature layers of a pretrained VGG-16 classification network. The gram matrix represents the style similarity between stylized results and real paintings in high-dimensional feature spaces. We, respectively, feed synthesis images and paintings into this module to calculate the distance and freeze the parameters of the VGG-16 network when updating the weight gradient in the back propagation phase:

$$\mathcal{L}_{sp} = \sum_{i=1}^4 \|VGG_{feat}^i(y_d) - VGG_{feat}^i(\tilde{y}_d)\|_1, \quad (3)$$

where $VGG_{feat}^i(\cdot)$ indicates the gram matrix of the i -th feature map of the VGG network pretrained on ImageNet.

Furthermore, we introduce a *conditional identity loss* to preserve content fidelity without affecting the output quality. Specifically, we constrain an identity mapping when using the same style image as style and content input:

$$\mathcal{L}_{cid} = \|Dec_c(E_c(y_d), E_s(y_d, d), d) - y_d\|_1, \quad (4)$$

where E_c and E_s encode features from the content image and the style image, respectively. Conditioned by the style label d , Dec_c reconstructs the style image y_d under the L1 metric, which obtains more realistic results compared with the L2 norm.

To generate realistic results, we use multi-scale patch-based discriminators for adversarial training and auxiliary classifiers for domain classification in Figure 1(d), similar to Reference [5]:

$$\begin{aligned}\mathcal{L}_{c_adv} = & \mathbb{E}_{y_d} [\log D_c(y_d)] \\ & + \mathbb{E}_{x, y_d, d} [\log(1 - D_c(G(x, y_d, d)))],\end{aligned}\quad (5)$$

where D_c is the multiscale discriminator network. The LSGAN [30] loss is used for adversarial training, and the auxiliary classification loss is applied to constrain the output stylized image $\tilde{y}_d$ and input style guidance y_d into the same style domain:

$$\begin{aligned}\mathcal{L}_{cls} = & \mathbb{E}_{y_d} [\mathcal{L}_{D_{cls}}(d|y_d)] \\ & + \mathbb{E}_{x, y_d, d} [\mathcal{L}_{D_{cls}}(d|G(x, y_d, d))],\end{aligned}\quad (6)$$

where $\mathcal{L}_{D_{cls}}$ is the cross-entropy loss.

The final objective for our generator, discriminator, style alignment module and auxiliary classifier is formulated as

$$\begin{aligned}\min_{E_s, G, D_{cls}} \max_{D_s, D_c} \mathcal{L}_{full} = & \lambda_1 \mathcal{L}_{c_adv} + \lambda_2 \mathcal{L}_{s_adv} \\ & + \lambda_3 (\mathcal{L}_{cid} + \mathcal{L}_{cp}) + \lambda_4 \mathcal{L}_{sp} + \lambda_5 \mathcal{L}_{cls},\end{aligned}\quad (7)$$

where $\{\lambda_i\}$ denotes the relative importance among these objectives. By several contrasting tests on these trade-off parameters, we determine to set $\lambda_1, \lambda_5 = 1, \lambda_2 = 10, \lambda_3 = 100, \lambda_4 = 30$ for achieving satisfactory visual performance.

4 EXPERIMENTS

4.1 Experimental Setup

Implementation details. We implement the proposed framework using PyTorch. The input resolution to our network is 256×256 . We set the dimension of style code to 20. For network training, we use Gaussian weight initialization and the Adam [15] optimizer. The learning rate, β_1 and β_2 are set to $2e - 4, 0.5$, and 0.999 , respectively. The model sets batch size as 1 and runs in the Titan XP device. The full model is trained with 350,000 iterations. One photo with one painting randomly selected is used to calculate related losses in training phase. We evaluate the results with extra data for testing.

Training and test data. For multi-domain training, we collect a total of 1,303 paintings from five artists, including Monet (458), Van Gogh (184), Cézanne (257), Gauguin (245), and Picasso (159) on Wikiart¹ as style reference images. Each artist corresponds to one style domain. The content image sets for training are from CycleGAN's photo2art dataset, which has a total number of 6,287 images. The test images are collected from Pexels² using the keywords *landscape* and *nature*. We prepare 974 natural photos in total as input content images.

Baseline methods. To demonstrate the controllability and diversity of our method, we compare it with seven recent style transfer methods (i.e., Gatys [11], AdaIN [12], WCT [22], AvatarNet [36], AAMS [41], SANet [33], and LinearST [21]), as well as seven I2I translation methods (i.e., CycleGAN [45], MUNIT [13], DRIT [19], StarGAN [5], UFDN [23], DMIT [42], and SMIT [34]). For a fair comparison, we use author-released source codes whenever possible and train all methods with default configurations on the same training set with same number of iterations, except CycleGAN. We train CycleGAN with dropout layers of probability 0.5 to make it feasible for multimodal image

¹<https://www.wikiart.org/>.

²<https://www.pexels.com/>.



Fig. 5. Exemplar-guided stylization results from different methods. The content and style images are shown in the left two columns. The remaining columns demonstrate output images generated by several popular style transfer approaches and our method.

generation. The new model is denoted as *CycleGAN_D*. We evaluate the performance of all models on the same *Pexels* test set as mentioned above.

4.2 Qualitative Evaluation

Qualitative comparison. Figure 5 shows the comparison of our exemplar-guided style transfer results with the ones of other approaches. Each row corresponds to one artist's style (domain) and different columns represent different methods. The corresponding content and style images are shown in the leftmost columns. Overall, our method achieves more visually plausible results than others. For example, Gatys et al. (Figure 5(c)) fails to preserve content semantic information well and reproduce sky in the first and third content images (first and third rows). AdaIN (Figure 5(d)) achieves high fidelity w.r.t input content images, but the results are often overblurred. WCT (Figure 5(e)) cannot obtain satisfactory results with severely distorted contents and less consistent style w.r.t style exemplars. AvatarNet (Figure 5(f)) and AAMS (Figure 5(i)) tend to generate either blurry results or images with granular artifacts. MUNIT (Figure 5(g)) performs better than WCT but also suffers from blurring issues and dirty appearance artifacts (e.g., second and third rows). DMIT (Figure 5(h)) cannot maintain a consistent style with the exemplar (second row). Finally, although SANet (Figure 5(j)) and LinearST (Figure 5(k)) present balanced appearance between content and style, they still suffer from content distortions (Figure 5(j), second row) and color deviations (Figure 5(k), fourth row). Compared with these approaches, our method has less



Fig. 6. Random sampling guided stylization results of different domains (artists). Random style codes are sampled from a standard normal distribution. (a) Content images, (b–d) Our stylization results using the styles of three artists.

artifacts while achieving better visual quality, i.e., content similarity and style consistency are well preserved.

Multimodal generation. Our method can generate diverse multimodal results in different ways for style guidance. To demonstrate this advantage, we generate multiple stylized images from the same content image for each domain representing the style of an artist (1) by random sampling in our learned style embedding space (see Figure 6) and (2) by using different reference images of the corresponding artist (see Figure 7). In both cases, our method can generate vivid stylized images that are consistent with the unique style of each artist. For example, the results guided by Picasso’s artwork are composed of large color blocks in an abstract style, whereas the ones guided by Monet’s work appear to be vague and are full of subtle strokes.

Multi-domain generation. Our style space decouples multi-domain control from multimodal generation. Figure 8 shows style transfer results with fixed style code and different artist domain labels. In general, our method can generate images of different styles that preserve the unique brush strokes and customary color collocations of each artist. For instance, the stylized photos from Picasso are more vivid and abstract, whereas the ones from Van Gogh generates many tiny strokes.

Style interpolation. To validate the smoothness of our latent style space, we present interpolation results using different guidance methods. Figure 9 shows image sequences generated by linear interpolation of two randomly sampled style codes in latent space. We can observe smooth and plausible style change as the interpolation weight varies. Figure 10 demonstrates interpolation results between multiple styles defined using the reference images shown in the four corners. We obtain satisfactory intra-domain (vertically) and inter-domain (horizontally) interpolation results.

4.3 Quantitative Evaluation

To evaluate our stylized image quality, we conduct quantitative comparison using two different metrics, i.e., LPIPS score and IS. The LPIPS metric [43] is defined as the perceptual distance between image pairs and is calculated as a weighted difference between their embeddings on a pretrained VGG16 network, where the weights are fit so that the metric agrees with human perceptual similarity judgments. IS is the expectation of KL-divergence distance between two



Fig. 7. Style transfer results using artworks of different artists as exemplars. The first column shows input content images and the first row shows the guiding exemplars. For each artist (domain), two exemplars are used to demonstrate intra-domain discrepancy.

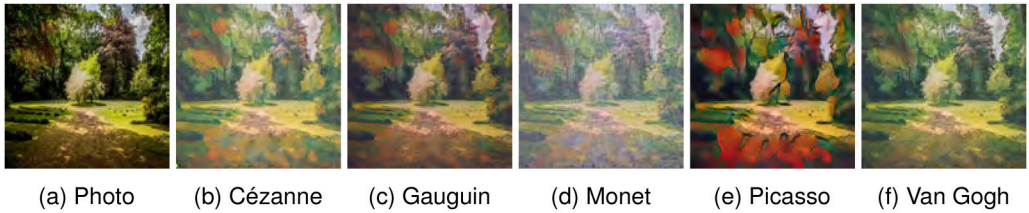


Fig. 8. Results of changing artist domains. We fix the content image and style code to conduct style transfer with different artists' labels.

sets of generated images in a feature space extracted by a pretrained Inception-V3 [38] network, which measures generation quality and diversity. To measure the quality for each artist style, we fine-tune the specific embedding network separately for each domain on our training data. We select 100 photos from the test set and stylize them in specific domains with 100 randomly sampled style codes for the IS metric and 19 different style code pairs for the LPIPS metric.

We report the mean and standard variation of LPIPS and IS in Table 2. As indicated by the scores, our method outperforms other methods in most test cases. Specifically, *CycleGAN_D* obtains the lowest LPIPS score and cannot generate diverse enough results, because the dropout layer only provides stochastic noise, which leads to limited changes in the output. StarGAN obtains the lowest IS score, which indicates that it cannot generate high-quality stylized images. We argue that the structure of StarGAN is designed for general multi-domain image translation and is not specifically optimized for the style transfer task. UFDN also does not perform well in most cases and fails to decouple content and style. Without any specific design and constraints, different style information is difficult to pass to a single generator to synthesize plausible stylization results. DRIT and MUNIT demonstrate better results than StarGAN and *CycleGAN_D* in terms of the two metrics. DMIT achieves superior performance in the IS metric but fails in the diversity evaluation



Fig. 9. Linear interpolation results of two random styles. The content images are shown on the left, whereas two randomly sampled styles are shown in the second leftmost and the rightmost columns.

Table 2. Comparisons of IS and LPIPS Scores of Different Stylization Methods for Each Painter Domain

Method	Photo2Monet (P2M)		Photo2Van Gogh (P2V)		Photo2Cézanne (P2C)		Photo2Gauguin (P2G)		Photo2Picasso (P2P)	
	IS	LPIPS	IS	LPIPS	IS	LPIPS	IS	LPIPS	IS	LPIPS
CycleGAN_D [45]	3.77 ± 0.06	0.16 ± 0.05	2.09 ± 0.02	0.16 ± 0.05	1.71 ± 0.03	0.18 ± 0.05	2.04 ± 0.03	0.19 ± 0.04	1.70 ± 0.02	0.23 ± 0.05
StarGAN [5]	2.64 ± 0.26	—	1.72 ± 0.03	—	1.51 ± 0.00	—	1.46 ± 0.01	—	1.22 ± 0.01	—
UFDN [23]	2.35 ± 0.02	—	1.93 ± 0.03	—	1.36 ± 0.01	—	1.49 ± 0.01	—	1.33 ± 0.01	—
DRIT [19]	3.50 ± 0.06	0.30 ± 0.12	1.97 ± 0.03	0.32 ± 0.12	1.50 ± 0.02	0.23 ± 0.09	1.97 ± 0.03	0.24 ± 0.10	1.56 ± 0.01	0.25 ± 0.10
MUNIT [13]	4.25 ± 0.10	0.44 ± 0.12	2.23 ± 0.03	0.46 ± 0.13	1.67 ± 0.02	0.37 ± 0.11	2.01 ± 0.03	0.39 ± 0.11	1.81 ± 0.02	0.54 ± 0.12
DMIT [42]	3.47 ± 0.06	0.26 ± 0.14	2.14 ± 0.04	0.26 ± 0.14	1.64 ± 0.03	0.26 ± 0.13	2.29 ± 0.05	0.26 ± 0.14	1.58 ± 0.01	0.25 ± 0.12
SMIT [34]	4.70 ± 0.01	0.39 ± 0.09	1.96 ± 0.03	0.36 ± 0.06	1.57 ± 0.02	0.36 ± 0.06	1.92 ± 0.03	0.39 ± 0.07	1.61 ± 0.02	0.38 ± 0.07
Ours+Exemplar	2.87 ± 0.04	0.21 ± 0.11	2.11 ± 0.02	0.30 ± 0.13	1.50 ± 0.02	0.19 ± 0.09	2.22 ± 0.03	0.20 ± 0.10	1.68 ± 0.04	0.31 ± 0.12
Ours+Sample	3.62 ± 0.04	0.48 ± 0.19	2.30 ± 0.04	0.50 ± 0.19	1.50 ± 0.03	0.43 ± 0.17	2.39 ± 0.04	0.45 ± 0.17	1.88 ± 0.03	0.49 ± 0.17
Real	7.01 ± 1.58	—	2.53 ± 0.61	—	3.77 ± 1.23	—	2.58 ± 0.44	—	2.92 ± 0.85	—

One hundred content images are, respectively, stylized 100 times at random to calculate the IS. The LPIPS metric is measured using 1,900 pairs of stylized images for the domain of each artist. For both metrics, the higher, the better. The last row presents the results of real data.

because of the usage of KL divergence. SMIT performs well on a single domain, but it is still not good enough in terms of overall quality and generation diversity. Our method outperforms these methods in most cases. The evaluation scores also show that our random sample-based generation is consistently better than exemplar-guided generation, indicating that our random-sample approach can synthesize diverse and high-quality stylized images. We assume that the exemplar-guidance synthesis is constrained by limited artworks, which leads to the shrinking of style space to several repeated points and the drop in performance. After removing the restriction, the random-sample approach outperforms other methods, thus proving the effectiveness of our framework.

Finally, to demonstrate that our method preserves consistent domain information for multi-modal generation, we conduct an experiment to reclassify stylized images into their corresponding style domains. We train a classification network using style labels from our training data. We compare our exemplar-guided and random sample-based approaches with two unified methods (i.e., StarGAN [13] and UFDN [23]) and report the corresponding results in Table 3. In most cases (four of five), our method leads to a higher classification accuracy, which indicates that our output distribution is closer to the reference style.

Table 3. Artist Classification Accuracy of Methods with a Unified Generator

Method	P2M	P2V	P2C	P2G	P2P	Overall
StarGAN [5]	0.46	0.88	0.14	0.15	0.11	0.348
UFDN [23]	0.92	0.56	0.71	0.16	0.10	0.490
Ours + Sample	0.76	0.26	0.41	0.33	0.31	0.414
Ours + Exemplar	0.97	0.47	0.83	0.69	0.50	0.692

For each domain, 100 test images from *Pexels.com* are used to perform stylization. The results are categorized by a ResNet-18 model fine-tuned on the test set.



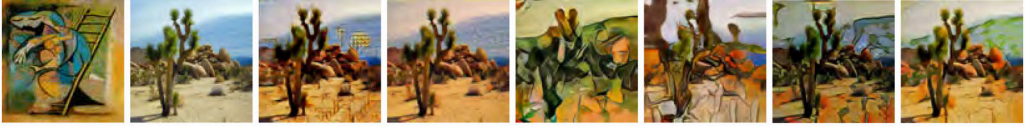
Fig. 10. Linear interpolation results of multiple styles. The input style images are shown in the four corners.

4.4 Ablation Study

To analyze the effect of each component in our framework, we conduct a series of ablation studies with certain components turned off and report the IS and LPIPS scores for evaluation. Table 4 shows the result of the ablation study. The first row shows the effect of replacing the multi-scale D_c with a standard discriminator. As shown, the corresponding model does not perform well either in terms of quality or diversity. The result without $\mathcal{L}_{sp}$ shows a considerable decrease in LPIPS score, because style consistency is no longer required. Without $\mathcal{L}_{cp}$, the network no longer preserves the content of input reference, thus performing poorly on IS and LPIPS scores. When $\mathcal{L}_{cid}$ is turned off, we observe a small increase in LPIPS score, which indicates a slightly improved diversity. However, the image quality decreases as indicated by a decreased IS score. Finally, we show the performance of the model replacing the **style alignment module (SAM)** with KL-divergence. On both indicators, the results are lower than our full model, indicating the effectiveness of our proposed module. Figure 11 demonstrates the visual quality under different ablation setups. Our full model in Figure 11(h) achieves the best trade-off between content fidelity and style consistency.

4.5 User Study

To evaluate our method quantitatively, we conduct user study to measure the preference of different stylization methods. We select six exemplar-guided style transfer methods, including WCT [22], AdaIN [12], MUNIT [13], DMIT [42], AAMS [41], and SANet [33], as baselines for



(a) Style (b) Photo (c) w/o D_c (d) w/o $\mathcal{L}_{cid}$ (e) w/o $\mathcal{L}_{cp}$ (f) w/o $\mathcal{L}_{sp}$ (g) w/o SAM (h) full model

Fig. 11. Qualitative results of the ablation study. The influencing factors are, respectively, removed and the corresponding results are compared with our full model.

Table 4. Ablation Study of Our Method

Method	IS	LPIPS
Ours w/o ms D_c	2.17 ± 0.03	0.23 ± 0.03
Ours w/o $\mathcal{L}_{sp}$	2.10 ± 0.03	0.13 ± 0.02
Ours w/o $\mathcal{L}_{cp}$	1.87 ± 0.02	0.11 ± 0.02
Ours w/o $\mathcal{L}_{cid}$	1.93 ± 0.03	0.65 ± 0.02
Ours w/o SAM	2.26 ± 0.03	0.42 ± 0.16
Ours full model	2.34 ± 0.04	0.47 ± 0.03

Different factors are removed separately to evaluate the corresponding effect on the results.

Table 5. Results of The user Study

Methods	Mean of win rate	
	Content fidelity	Style preference
Ours vs WCT [22]	82.08%	57.66%
Ours vs AdaIN [12]	68.56%	56.70%
Ours vs MUNIT [13]	81.73%	68.53%
Ours vs DMIT [42]	51.68%	58.39%
Ours vs AAMS [41]	83.21%	75.57%
Ours vs SANet [33]	73.08%	54.10%

In each row, we report the average percentages that our results are selected when compare to these from the corresponding baseline methods.

comparison. We hire 30 annotators of different backgrounds and from different regions to answer 65 randomly generated questions. Given the content image and the exemplar, we show the subjects two stylization results (one by our method and the other from a baseline approach) in random order. For each pair of stylized images, the annotator is asked to answer two questions: (1) which one has higher fidelity to the content image; and (2) which one has the preferred style. The results of our user study are reported in Table 5. In each row, we provide the average percentages, which means the proportion of our results being selected when compared with baseline methods by the two aforementioned questions. As shown in Table 5, our results are more preferred than those from baseline approaches by both measures.

5 CONCLUSION

In this study, we present a new framework to achieve multimodal and multi-domain style transfer. To enable image stylization via exemplar-based and randomly sampled guidance, we propose a novel style alignment module to construct an embedding style space. The constructed space eliminates the explicit distribution gaps among various style domains and enables image-guided

style feature extraction and random style code generation, while reducing the risk of mode collapse due to the improper constraint by using KL-divergence loss. Our framework shows superior performance in multimodal and multi-domain style transfer. Extensive qualitative and quantitative evaluations demonstrate that our method outperforms previous style transfer methods.

REFERENCES

- [1] Asha Anooosheh, Eiríkur Agustsson, Radu Timofte, and Luc Van Gool. 2018. ComboGAN: Unrestrained scalability for image domain translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'18)*. IEEE, Los Alamitos, CA, 896–8967. <https://doi.org/10.1109/CVPRW.2018.00122>
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. Retrieved from <https://arXiv:stat.ML/1607.06450>.
- [3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. 2017. CVAE-GAN: Fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. IEEE, Los Alamitos, CA, 2764–2773. <https://doi.org/10.1109/ICCV.2017.299>
- [4] Simyung Chang, SeongUk Park, John Yang, and Nojun Kwak. 2019. Sym-parameterized dynamic inference for mixed-domain image translation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'19)*. IEEE, Los Alamitos, CA, 4803–4811. <https://doi.org/10.1109/ICCV.2019.00490>
- [5] Yunje Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. IEEE, Los Alamitos, CA, 8789–8797. <https://doi.org/10.1109/CVPR.2018.00916>
- [6] Yunje Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'20)*. IEEE, Los Alamitos, CA, 8185–8194. <https://doi.org/10.1109/CVPR42600.2020.00821>
- [7] Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Ma Chongyang, and Changsheng Xu. 2021. Arbitrary video style transfer via multi-channel correlation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI'21)*.
- [8] Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. 2020. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the 28th ACM International Conference on Multimedia (MM'20)*. ACM, New York, NY, 27192727. <https://doi.org/10.1145/3394171.3414015>
- [9] Lars Doyle, Forest Anderson, Ehren Choy, and David Mould. 2019. Automated pebble mosaic stylization of images. *Comput. Visual Media* 5, 1 (2019), 33–44. <https://doi.org/10.1007/s41095-019-0129-0>
- [10] Alexei A. Efros and William T. Freeman. 2001. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'01)*. ACM, New York, NY, 341–346. <https://doi.org/10.1145/383259.383296>
- [11] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. IEEE, Los Alamitos, CA, 2414–2423. <https://doi.org/10.1109/CVPR.2016.265>
- [12] X. Huang and S. Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. IEEE, Los Alamitos, CA, 1510–1519. <https://doi.org/10.1109/ICCV.2017.167>
- [13] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. Springer, Cham, Switzerland, 179–196. https://doi.org/10.1007/978-3-030-01219-9_11
- [14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. IEEE, Los Alamitos, CA, 5967–5976. <https://doi.org/10.1109/CVPR.2017.632>
- [15] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. Retrieved from <https://arXiv:cs.LG/1412.6980>.
- [16] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. Retrieved from <https://arXiv:stat.ML/1312.6114>.
- [17] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. 2019. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'19)*. IEEE, Los Alamitos, CA, 4421–4430. <https://doi.org/10.1109/ICCV.2019.00452>
- [18] Dmytro Kotovenko, Artsiom Sanakoyeu, Pingchuan Ma, Sabine Lang, and Björn Ommer. 2019. A content transformation block for image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19)*. IEEE, Los Alamitos, CA, 10024–10033. <https://doi.org/10.1109/CVPR.2019.01027>

- [19] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. Springer, Cham, Switzerland, 36–52. https://doi.org/10.1007/978-3-030-01246-5_3
- [20] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. 2020. DRIT++: diverse image-to-image translation via disentangled representations. *Int. J. Comput. Vision* 128 (2020), 24022417. <https://doi.org/10.1007/s11263-019-01284-z>
- [21] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. 2019. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19)*. IEEE, Los Alamitos, CA, 3804–3812. <https://doi.org/10.1109/CVPR.2019.00393>
- [22] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates, Inc., Red Hook, NY, 385395. <https://doi.org/10.5555/3294771.3294808>
- [23] Alexander H. Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. 2018. A unified feature disentangler for multi-domain image translation and manipulation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Curran Associates, Red Hook, NY, 2590–2599. <http://papers.nips.cc/paper/7525-a-unified-feature-disentangler-for-multi-domain-image-translation-and-manipulation.pdf>.
- [24] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates, Red Hook, NY, 700–708. <https://doi.org/10.5555/3294771.3294838>
- [25] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. 2019. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'19)*. IEEE, Los Alamitos, CA, 10550–10559. <https://doi.org/10.1109/ICCV.2019.01065>
- [26] Yahui Liu, Marco De Nadai, Jian Yao, Nicu Sebe, Bruno Lepri, and Xavier Alameda-Pineda. 2020. GMM-UNIT: Unsupervised Multi-Domain and Multi-Modal Image-to-Image Translation via Attribute Gaussian Mixture Modeling. Retrieved from <https://arXiv.cs.CV/2003.06788>.
- [27] Liqian Ma, Xu Jia, Stamatios Georgioulis, Tinne Tuytelaars, and Luc Van Gool. 2019. Exemplar guided unsupervised image-to-image translation with semantic consistency. In *Proceedings of the 7th International Conference on Learning Representations (ICLR'19)*. Retrieved from <https://openreview.net/forum?id=SlITg3RqYQ>.
- [28] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2016. Adversarial Autoencoders. Retrieved from <https://arXiv.cs.LG/1511.05644>.
- [29] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. 2019. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19)*. IEEE, Los Alamitos, CA, 1429–1437. <https://doi.org/10.1109/CVPR.2019.00152>
- [30] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. IEEE, Los Alamitos, CA, 2813–2821. <https://doi.org/10.1109/ICCV.2017.304>
- [31] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. Retrieved from <https://arXiv.cs.LG/1411.1784>.
- [32] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, Vol. 70. JMLR, Cambridge, MA, 2642–2651. <https://doi.org/10.5555/3305890.3305954>
- [33] Dae Young Park and Kwang Hee Lee. 2019. Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19)*. IEEE, Los Alamitos, CA, 5873–5881. <https://doi.org/10.1109/CVPR.2019.00603>
- [34] Andrés Romero, Pablo Arbeláez, Luc Van Gool, and Radu Timofte. 2019. SMIT: Stochastic multi-label image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'19)*. IEEE, Los Alamitos, CA, 3285–3294. <https://doi.org/10.1109/ICCVW.2019.00410>
- [35] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Björn Ommer. 2018. A style-aware content loss for real-time HD style transfer. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. Springer, Cham, Switzerland, 715–731. https://doi.org/10.1007/978-3-030-01237-3_43
- [36] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. 2018. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. IEEE, Los Alamitos, CA, 8242–8250. <https://doi.org/10.1109/CVPR.2018.00860>
- [37] Kihyuk Sohn, Xinchun Yan, and Honglak Lee. 2015. Learning structured output representation using deep conditional generative models. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15)*, Vol. 2. MIT Press, Cambridge, MA, 34833491. <https://doi.org/10.5555/2969442.2969628>

- [38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. IEEE, Los Alamitos, CA, 2818–2826. <https://doi.org/10.1109/CVPR.2016.308>
- [39] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2017. Instance Normalization: The Missing Ingredient for Fast Stylization. Retrieved from <https://arXiv:cs.CV/1607.08022>.
- [40] Fu-En Yang, Jing-Cheng Chang, Chung-Chi Tsai, and Yu-Chiang Frank Wang. 2020. A multi-domain and multi-modal representation disentangler for cross-domain image manipulation and classification. *IEEE Trans. Image Process.* 29 (2020), 2795–2807. <https://doi.org/10.1109/TIP.2019.2952707>
- [41] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. 2019. Attention-aware multi-stroke style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19)*. IEEE, Los Alamitos, CA, 1467–1475. <https://doi.org/10.1109/CVPR.2019.00156>
- [42] Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas Li, and Ge Li. 2019. Multi-mapping Image-to-image translation via learning disentanglement. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS'19)*. Curran Associates, Red Hook, NY, 2994–3004. Retrieved from <http://papers.neurips.cc/paper/8564-multi-mapping-image-to-image-translation-via-learning-disentanglement.pdf>.
- [43] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. IEEE, Los Alamitos, CA, 586–595. <https://doi.org/10.1109/CVPR.2018.00068>
- [44] Shuyang Zhang, Runze Liang, and Miao Wang. 2019. ShadowGAN: Shadow synthesis for virtual objects with conditional adversarial networks. *Comput. Visual Media* 5, 1 (2019), 105–115. <https://doi.org/10.1007/s41095-019-0136-1>
- [45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. IEEE, Los Alamitos, CA, 2242–2251. <https://doi.org/10.1109/ICCV.2017.244>

Received September 2020; revised January 2021; accepted February 2021