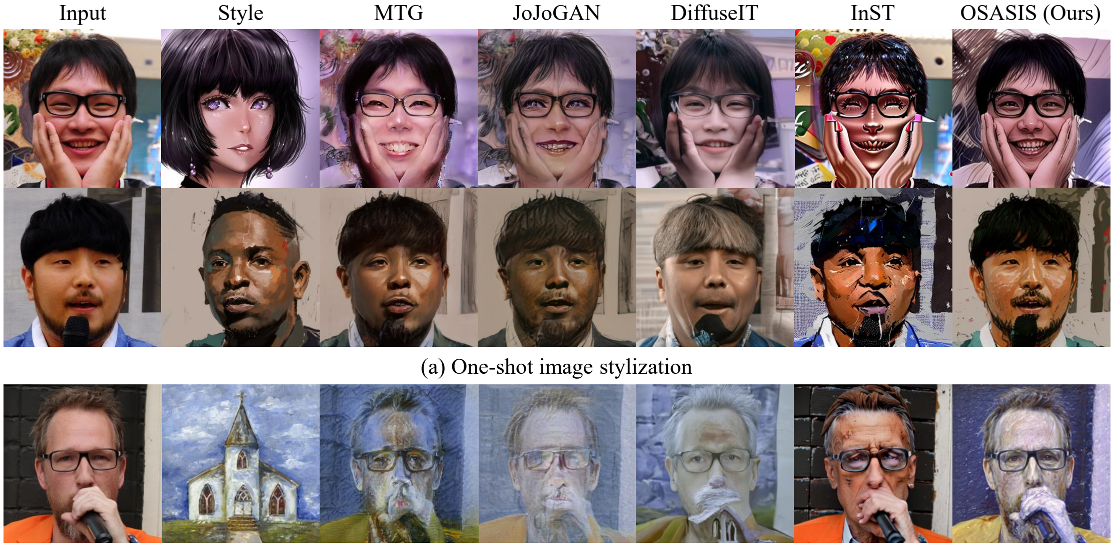
One-Shot Structure-Aware Stylized Image Synthesis

Hansam Cho[1,2*], Jonghyun Lee[1,2*], Seunggyu Chang[1], Yonghyun Jeong[1†]
[1]NAVER Cloud, [2]School of Industrial and Management Engineering, Korea University
{chosam95, tomtom1103}@korea.ac.kr,
{seunggyu.chang, yonghyun.jeong}@navercorp.com



(a) One-shot image stylization

(b) One-shot image stylization with OOD reference image

Figure 1. OSASIS is able to (a) stylize an input image with a single reference image while robustly preserving the structure and content of the input image. It is also able to (b) incorporate out-of-domain (OOD) data as the reference image while other baseline methods fail

Abstract

While GAN-based models have been successful in image stylization tasks, they often struggle with structure preservation while stylizing a wide range of input images. Recently, diffusion models have been adopted for image stylization but still lack the capability to maintain the original quality of input images. Building on this, we propose OSASIS: a novel one-shot stylization method that is robust in structure preservation. We show that OSASIS is able to effectively disentangle the semantics from the structure of an image, allowing it to control the level of content and style implemented to a given input. We apply OSASIS to various experimental settings, including stylization with out-of-domain reference images and stylization with text-driven manipulation. Results show that OSASIS outperforms other stylization methods, especially for input images that were rarely encountered during training, providing a promising solution to stylization via diffusion models. The source code can be found at https://github.com/hansam95/OSASIS.

1. Introduction

In the literature of generative models, image stylization refers to training a model in order to transfer the style of a reference image to various input images during inference [21, 22, 28]. However, collecting a sufficient number of images that share a particular style for training can be difficult. Consequently, one-shot stylization has emerged as a viable and practical solution, with generative adversarial networks (GANs) showing promising results [2, 16, 36, 38, 39].

Despite significant advancements in GAN-based stylization techniques, the accurate preservation of an input image's structure continues to pose a significant challenge. This difficulty is particularly pronounced for input images that contain elements infrequently encountered during training, often characterized by complex structural nuances that diverge from those observed in more commonly presented images. Figure 1(a) illustrates this challenge, where entities such as hands and microphones, when processed through GAN-based stylization, diverge considerably from their original structural integrity. In addition, GAN-based styl-

[*] Work done during an internship at NAVER Cloud.
[†] Corresponding author.

ization methods often fail to accurately separate the structure and style of the reference image during inference. As shown in Figure 1(b), the lack of disentanglement results in structural artifacts from reference images bleeding into the stylized image.

Recently, diffusion models (DMs) have shown remarkable performance in various image-related tasks, including high-fidelity image generation [25, 26], super resolution [27], and text-driven image manipulation [13, 15, 17]. For image stylization, several studies, including DiffuseIT [15] and InST [37], have been proposed. However, they primarily focus on developing a diffusion model framework tailored to the stylization task. In contrast, our work prioritizes preserving the structure of the input image over solely introducing an appropriate diffusion model for stylization. As illustrated in Figure 1(a), it can be seen that the capability to preserve structure doesn't stem from diffusion models itself, but from our methodology.

In this study, we propose **O**ne-shot **S**tructure-**A**ware **S**tylized **I**mage **S**ynthesis (OSASIS), which effectively disentangles the structure and transferable semantics of a style image within the structure of a diffusion model. OSASIS selects an appropriate encoding timestep of a structural latent code to control the strength of structure preservation and enhances its preservation ability through a structure-preserving network. To acquire a semantically meaningful latent, we utilize the semantic encoder proposed in diffusion autoencoders (DiffAE) [23]. Following the approach of mind the gap (MTG) [39], we bridge the domain gap by finetuning a pretrained DM using a combination of directional CLIP losses. Once trained, we find that by properly conditioning the semantic latent code, our method achieves structure-aware image stylization.

We conduct qualitative and quantitative experiments on a wide range of input and style images. By quantitatively extracting data with rare structural elements from the training set (*i.e.* low-density images), we show that OSASIS is robust in structure preservation, outperforming other methods. In addition, we directly optimize the semantic latent code for text-driven manipulation. Combining the optimized latent with the finetuned DM, OSASIS is able to generate stylized images with manipulated attributes.

2. Background

2.1. Diffusion Models

Diffusion models are latent variable models that are trained to reverse a forward process[8]. The forward process, which is defined as a Markov chain with a Gaussian transition defined in Eq.1, involves iteratively mapping an image to a predefined prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ over T steps. DDPM[8] proposes to parameterize the reverse process defined in Eq. 2 with a noise prediction network $\epsilon_\theta(\mathbf{x}_t, t)$, which is trained with the loss function $\mathcal{L}_{\text{simple}}$ in Eq.3.

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = N(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = N(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t\mathbf{I}) \tag{2}$$

$$\text{where} \quad \mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{1-\beta_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(\mathbf{x}_t, t))$$

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon_\theta(\mathbf{x}_t, t) - \epsilon)\|^2\right], \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{3}$$

In contrast to DDPM, DDIM [29] defines the forward process as non-Markovian and derives the corresponding reverse process as Eq. 5, in which $f_\theta(x_t, t)$ is the model's prediction of $\mathbf{x}_0$. DDIM also introduces an image encoding method by deriving ordinary differential equations (ODEs) corresponding with the reverse process. By reversing the ODE, DDIM introduces an image encoding process, defined as Eq. 4.

$$\mathbf{x}_{t+1} = \sqrt{\alpha_{t+1}} f_\theta(\mathbf{x}_t, t) + \sqrt{1-\alpha_{t+1}}\epsilon_\theta(\mathbf{x}_t, t) \tag{4}$$

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}} f_\theta(\mathbf{x}_t, t) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(\mathbf{x}_t, t) \tag{5}$$

$$\text{where} \quad f_\theta(\mathbf{x}_t, t) = \frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{\alpha_t}}$$

For our work, we adopt specific terminologies to refer to the forward and reverse process of DDPM, in which we call forward DDPM and reverse DDPM, respectively. Similarly, the denoising reverse process of DDIM is referred to as reverse DDIM, while the deterministic image encoding process of DDIM is referred to as forward DDIM.

2.2. Diffusion Autoencoders

DiffAE [23] proposes a semantic encoder Enc_ϕ that encodes a given image $\mathbf{x}_0$ to a semantically rich latent variable $\mathbf{z}_{\text{sem}}$, represented as:

$$\mathbf{z}_{\text{sem}} = \text{Enc}_\phi(\mathbf{x}_0). \tag{6}$$

This latent variable has been shown to be linear, decodable, and semantically meaningful, hence being an attractive property that our model seeks to leverage. Similar to the aforementioned forward DDIM, DiffAE is also able to encode an image given $\mathbf{z}_{\text{sem}}$ to a fully reconstructable latent $\mathbf{x}_T$ by Eq. 7, in which we denote forward DiffAE. Correspondingly, the forward DiffAE encoded latent $\mathbf{x}_T$ can be decoded by conditioning itself and $\mathbf{z}_{\text{sem}}$ to the reverse process defined as Eq. 8, referred to as reverse DiffAE.

$$\mathbf{x}_{t+1} = \sqrt{\alpha_{t+1}} f_\theta(\mathbf{x}_t, t, \mathbf{z}_{\text{sem}}) + \sqrt{1-\alpha_{t+1}}\epsilon_\theta(\mathbf{x}_t, t, \mathbf{z}_{\text{sem}}) \tag{7}$$

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}} f_\theta(\mathbf{x}_t, t, \mathbf{z}_{\text{sem}}) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(\mathbf{x}_t, t, \mathbf{z}_{\text{sem}}) \tag{8}$$

$$\text{where} \quad f_\theta(\mathbf{x}_t, t, \mathbf{z}_{\text{sem}}) = \frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{x}_t, t, \mathbf{z}_{\text{sem}})}{\sqrt{\alpha_t}}$$

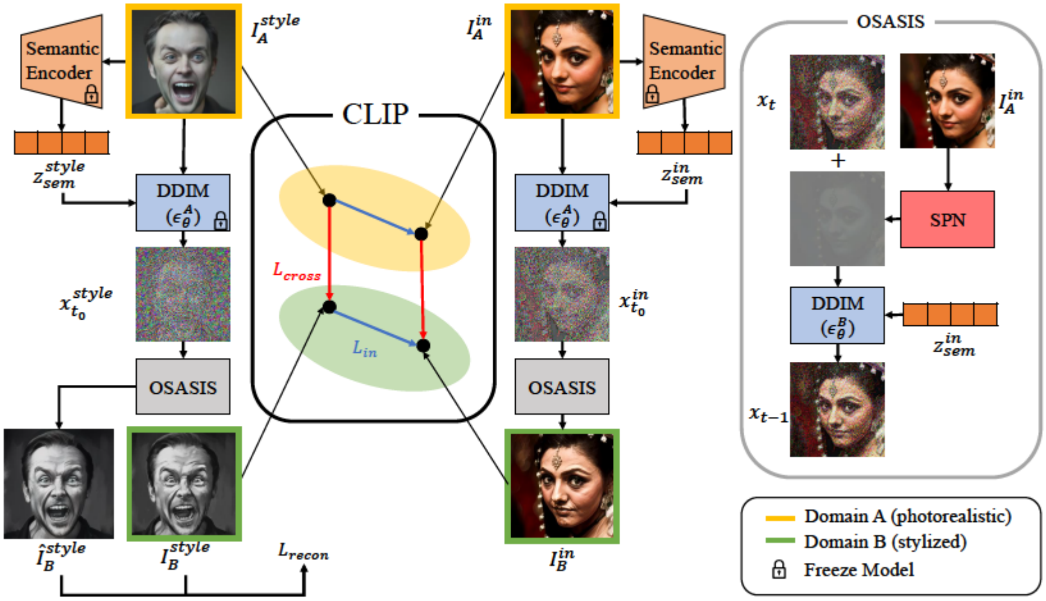


Figure 2. Overview of OSASIS. During finetuning, cross-domain loss compares the photorealistic image (bounded yellow) to its stylized counterparts (bounded green). Concurrently, the in-domain loss gauges the alignment of directional shifts within the same domain, which are delineated by yellow and green. Reconstruction loss compares the original style image with a reconstructed counterpart. Intuitively, the combination of the directional losses guarantees that for each iteration the generated I_B^{in} is positioned for projection vectors from I_B^{style} and I_A^{in} to be collinear to its cross-domain and in-domain counterparts in the CLIP space.

3. Methods

Our approach aims to achieve effective stylization by initially disentangling the structural and semantic information of images. We define structural information as the overall outline of an image, whereas we further deconstruct the semantics of an image into a combination of content and style. To initially disentangle the semantics from the structure, we employ two distinct latent codes: the structural latent code $\mathbf{x}_{\mathbf{t}_0}$ and the semantic latent code $\mathbf{z}_{\text{sem}}$. We finetune a pretrained DDIM ϵ_θ^A conditioned on the semantic latent code $\mathbf{z}_{\text{sem}}$ via CLIP directional losses in order to bridge the domain gap between the input and style images. Once finetuned, we control the amount of low-level visual features (*e.g.* texture and color), referred to as style, and high-level visual features (*e.g.* object and identity), referred to as content during inference. Proper conditioning of $\mathbf{z}_{\text{sem}}$ to the finetuned DDIM allows us to achieve this control, effectively performing stylization. Furthermore, we directly optimize the semantic latent code $\mathbf{z}_{\text{sem}}$ for text-driven manipulation. By combining the optimized latent with the finetuned DDIM, OSASIS is able to produce stylized images with manipulated attributes. Figure 2 provides an overview of our method.

3.1. Training

To ensure that the changes in the CLIP embedding space occur in the desired direction, we prepare a single image I_B^{style} from a stylized domain (denoted domain B) and aim to convert it to a photorealistic domain (denoted domain A). Recent studies have shown that pretrained DMs can generate domain-specific images based on unseen domain images [13, 20]. Building on this, we utilize a pretrained DDPM ϵ_θ to generate a single image I_A^{style} from domain A that is semantically aligned with I_B^{style}. I_B^{style} is encoded to a specific timestep t_0 by utilizing the forward DDPM:

$$\mathbf{x}_{\mathbf{t}_0} = \sqrt{\alpha_{\mathbf{t}_0}}\mathbf{x}_0 + \sqrt{1-\alpha_{\mathbf{t}_0}}\mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{9}$$

and subsequently from $\mathbf{x}_{\mathbf{t}_0}$, I_A^{style} is generated by following the reverse DDPM:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(\mathbf{x}_t, t)) + \sigma_t \mathbf{z}, \tag{10}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. After initializing I_A^{style} and I_B^{style}, we proceed to freeze a pretrained DDIM ϵ_θ^A and the semantic encoder Enc_ϕ proposed by DiffAE, which is utilized during the image encoding process. Additionally, we create a copy of ϵ_θ^A called ϵ_θ^B, which is finetuned through a combination of CLIP directional losses and a reconstruction loss. During training, we note that I_A^{in} is generated from the pretrained DiffAE. Consequently, our method enables training without the necessity for a dataset.

Structural Latent Code ϵ_θ^B optimizes towards generating I_B^{in} that reflects the semantics of the style image. However, since I_A^{style} and I_B^{style} stays fixed while I_A^{in} is constantly generated, it is crucial to carefully choose an encoding process that generates I_B^{in} that preserves the structural integrity of I_A^{in}. In order to achieve this, we utilize forward DiffAE to encode I_A^{in} by computing Eq. 7. First, I_A^{in} is encoded into a semantic latent code $\mathbf{z}_{\text{sem}}^{\text{in}} = \text{Enc}_\phi(I_A^{\text{in}})$. By following the forward process, $\mathbf{z}_{\text{sem}}^{\text{in}}$ is conditioned to the frozen DDIM $\boldsymbol{\epsilon}_\theta^A$,

$$\mathbf{x}_{t+1} = \sqrt{\alpha_{t+1}} f_\theta(\mathbf{x}_t, t, \mathbf{z}_{\text{sem}}^{\text{in}}) + \sqrt{1-\alpha_{t+1}} \epsilon_\theta^A(\mathbf{x}_t, t, \mathbf{z}_{\text{sem}}^{\text{in}}) \tag{11}$$

resulting in I_A^{in} encoded to a structural latent code $\mathbf{x}_{\mathbf{t}_0}^{\text{in}}$. The input image is encoded to a specific timestep $\mathbf{t}_0$ which can be adjusted to control the level of structure preservation.

Structure-Preserving Network Although the structural latent code $\mathbf{x}_{\mathbf{t}_0}^{\text{in}}$ succeeds in preserving the overall structure of the generated images I_A^{in}, the encoding process defined in Eq. 11 inherently adds noise, which inevitably results in the loss of structural information. To address this, we introduce a structure-preserving network (SPN), which utilizes a 1x1 convolution that effectively preserves the spatial information and structural integrity of I_A^{in}. To generate the output of the next timestep $\mathbf{x}_{t-1}$, we use reverse DiffAE with SPN:

$$\mathbf{x}_t^{SPN} = SPN(I_A^{\text{in}}) \tag{12}$$

$$\mathbf{x}_t' = \mathbf{x}_t + \lambda_{SPN} * \mathbf{x}_t^{SPN} \tag{13}$$

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}} f_\theta(\mathbf{x}_t', t, \mathbf{z}_{\text{sem}}^{\text{in}}) + \sqrt{1-\alpha_{t-1}} \epsilon_\theta^B(\mathbf{x}_t', t, \mathbf{z}_{\text{sem}}^{\text{in}}) \tag{14}$$

We add the output of the SPN to $\mathbf{x}_t$, *e.g.* the previous timestep output of $\boldsymbol{\epsilon}_\theta^B$, and feed it into our training target $\boldsymbol{\epsilon}_\theta^B$. We regulate the degree of spatial information reflected in the model by multiplying the output of the SPN $\mathbf{x}_t^{SPN}$ by λ_{SPN}. After fully reversing the timesteps, I_B^{in} is generated.

Loss Function Inspired by MTG [39], we train $\boldsymbol{\epsilon}_\theta^B$ by optimizing our total loss, which is comprised of the cross-domain loss, in-domain loss, and reconstruction loss. The cross-domain loss aims to align the direction of changes from domain A to domain B, ensuring that the change from I_A^{in} to I_B^{in} is kept consistent with the change from I_A^{style} to I_B^{style}. Although the cross-domain loss provides the changes in semantic information for the model to optimize upon, it often leads to unintended changes when implemented alone. Hence the in-domain loss is introduced to provide additional information, measuring the similarity of changes within both domains A and B.

The reconstruction loss provides additional guidance in capturing the cross-domain change from I_A^{style} to I_B^{style}. Similar to our encoding process of Eq. 11, I_A^{style} is encoded to a structural latent code $\mathbf{x}_{\mathbf{t}_0}^{\text{style}}$ conditioned on semantic latent code $\mathbf{z}_{\text{sem}}^{\text{style}}$. Following Eq. 12- 14, *i.e.* the process of generating the output of the next timestep conditioned on $\mathbf{z}_{\text{sem}}^{\text{style}}$, $\hat{I}_B^{\text{style}}$ is generated. The reconstruction loss is calculated by comparing $\hat{I}_B^{\text{style}}$ with I_B^{style}, comprised of the L_1 loss, perceptual similarity loss [35], and the L_1 CLIP embedding loss. Detailed information regarding the loss function and experimental setup is available in the supplementary material.

3.2. Sampling

Mixing Content and Style Once trained, the model $\boldsymbol{\epsilon}_\theta^B$ is capable of stylizing images from domain A to B. Stylizing an image involves mixing two images in its latent space. While this process is straightforward with StyleGAN [10], it is still an ongoing research area for diffusion models. Unlike the original DiffAE which conditions a single semantic latent code $\mathbf{z}_{\text{sem}}$ to the feature maps of DDIM, we discover that properly conditioning $\mathbf{z}_{\text{sem}}$ to the feature maps of $\boldsymbol{\epsilon}_\theta^B$ achieves content and style mixing. This is done by conditioning $\mathbf{z}_{\text{sem}}^{\text{style}}$ to its low-level feature maps to transfer the style of a style image, and conditioning $\mathbf{z}_{\text{sem}}^{\text{in}}$ to its high-level feature maps to transfer the content of an input image. The change point of conditioning is set as f_{ch}. Since $\boldsymbol{\epsilon}_\theta^B$ is a UNet-based model, conditioning the two latents is symmetrical. To preserve the structural information of the input image, we use $\mathbf{x}_{\mathbf{t}_0}^{\text{in}}$ as the structural latent code. The detailed process of sampling is described in the supplementary material.

Text-driven Manipulation Instead of optimizing a model, we directly optimize the semantic latent code of input image $\mathbf{z}_{\text{sem}}^{\text{in}}$ to achieve text-driven manipulation. Similar to previous works [13], we use CLIP directional loss for optimization. After optimization, the optimized $\mathbf{z}_{\text{sem}}^{\text{in}}$ can be passed into the $\boldsymbol{\epsilon}_\theta^B$to incorporate the style into the image. Comprehensive details about the experimental approach for text-driven manipulation are available in the supplementary material.

4. Experiments

For evaluating our approach, we focus on images from low-density regions that include rarely encountered attributes during training. This is ideal for demonstrating the structure-preserving ability of our method as low-density region images contain diverse objects and occlusions that obscure the subject. To select these images, we leverage the property that the encoded stochastic subcode $\mathbf{x}_T$ tends to show residuals of the original input image rather than being normally distributed, as shown by DiffAE [23]. We randomly select 20,000 images from the FFHQ dataset [10], which is the dataset $\boldsymbol{\epsilon}_\theta^A$ was trained on. We reconstruct each image using its semantic subcode $\mathbf{z}_{\text{sem}}$ and stochastic subcode $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ (*i.e.* stochastic reconstruction). We compare the reconstructed image with the original image using perceptual similarity loss [35] to determine the quality of the reconstruction. We hypothesize that high-density region images that contain frequently encountered attributes would be reconstructed accurately, whereas those from low-density regions would not. Figure 3 shows that our hypothe-

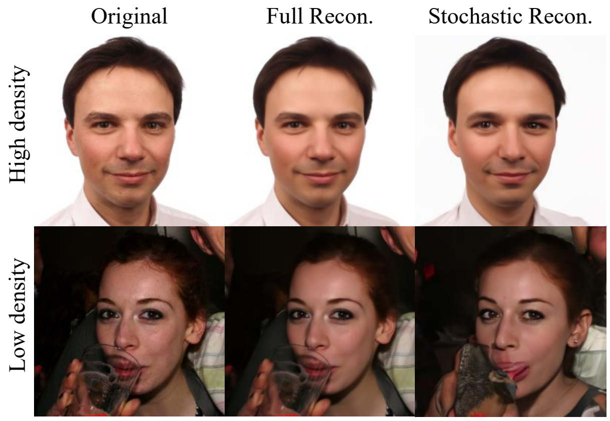


Figure 3. High and Low-density images. Full Recon. refers to reconstruction via conditioning its encoded $\mathbf{z}_{\text{sem}}$ and $\mathbf{x}_T$, whereas Stochastic Recon. refers to reconstruction via conditioning its encoded $\mathbf{z}_{\text{sem}}$ and $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

sis is well supported. Finally, we select the images from the top 100 highest LPIPS score group (*i.e.* low-density) and lowest LPIPS score group (*i.e.* high-density).

4.1. Qualitative Comparison

In Figure 4, we present a comparative analysis of the performance of OSASIS against other stylization methods. Our results demonstrate that OSASIS outperforms other methods in terms of preserving the overall structure while stylizing. MTG [39] and JoJoGAN [2] use outdated inversion methods that struggle to preserve the diverse structure of input images. Nonetheless, recent advancements in GAN-based inversion techniques have demonstrated significant improvements in handling out-of-distribution input images for editing purposes [24, 31, 33]. To ensure a fair comparison, we employ HFGI [31] for MTG and JoJoGAN. Despite these adjustments, OSASIS remains distinguished in its structural preservation capabilities. Furthermore, GAN-based methods produce unintended modifications, such as changes in facial expressions. Since DiffuseIT [15] stylizes images without training, they struggle to overcome the domain gap between the input and style images. InST [37] utilizes textual inversion to deduce the style image's concept and subsequently conditions its generation procedure on this concept. However, the guidance strategy outlined in [7] tends to produce style-concentrated images, leading to unintended variations such as changes in facial expressions and identities. Moreover, as noted in InST, there are difficulties in faithfully transferring the color of the style images. More qualitative comparison results are provided in the supplementary material.

4.2. Quantitative Comparison

We conduct a quantitative comparison with other methods. By using ArtFID [32] as a metric for effective stylization and the identity similarity measure with ArcFace [3] to assess content preservation. For measuring structure preservation, we employ the structure distance metric [30]. For our source of style images, we select five style images from each of three datasets: i) AAHQ [18], ii) MetFaces [11], iii) style images used in previous researches. In Table 1, we evaluate OSASIS against MTG [39], JoJoGAN [2], DiffuseIT [15], and InST [37]. For our initial pretrained $\boldsymbol{\epsilon}_\theta^A$ and all baseline models, we use publicly available pretrained models that were trained on FFHQ. As previously mentioned, HFGI [31] is employed to invert input images for MTG and JoJoGAN. Note that Table 1 presents outcomes only for low-density images, while a comprehensive comparison is presented in the supplementary material. Our results indicate that while MTG occasionally achieves a better ArtFID score than our model, we outperform significantly in terms of identity similarity and structure preservation. Additionally, while DiffuseIT excels at preserving the identity and the structure of the image, it exhibits inferior stylization results compared to GAN-based methods due to a domain gap between the input and style images.

Methods	ArtFID↓		
	AAHQ	MetFaces	Prev
MTG+HFGI	36.39	**38.02**	37.27
JoJoGAN+HFGI	40.41	44.74	41.09
DiffuseIT	44.93	53.35	48.18
InST	38.16	50.33	35.86
OSASIS(Ours)	**34.89**	43.20	**33.20**
Methods	ID Similarity↑		
	AAHQ	MetFaces	Prev
MTG+HFGI	0.3730	0.4656	0.4063
JoJoGAN+HFGI	0.5145	0.5207	0.4743
DiffuseIT	**0.6992**	0.7158	0.6994
InST	0.2253	0.2188	0.2238
OSASIS(Ours)	0.6825	**0.7323**	**0.7029**
Methods	Structure Distance↓		
	AAHQ	MetFaces	Prev
MTG+HFGI	0.0386	0.0350	0.0360
JoJoGAN+HFGI	0.0411	0.0454	0.0430
DiffuseIT	**0.0309**	0.0300	**0.0310**
InST	0.0492	0.0443	0.0488
OSASIS(Ours)	0.0361	**0.0295**	0.0391

Table 1. Quantitative comparison. ArtFID evaluates the pertinence of stylization, whereas ID similarity and structure distance measure whether the stylized image stays true to its original input.

4.3. OOD Reference Image

OSASIS is able to stylize images with out-of-domain (OOD) reference images, a feature that is not commonly seen in GAN-based methods. OSASIS can effectively disentangle semantics from the structure, resulting in only the style factor of the style image being transferred to the input image. In contrast, GAN-based methods have entangled style codes in terms of structure and semantics, which

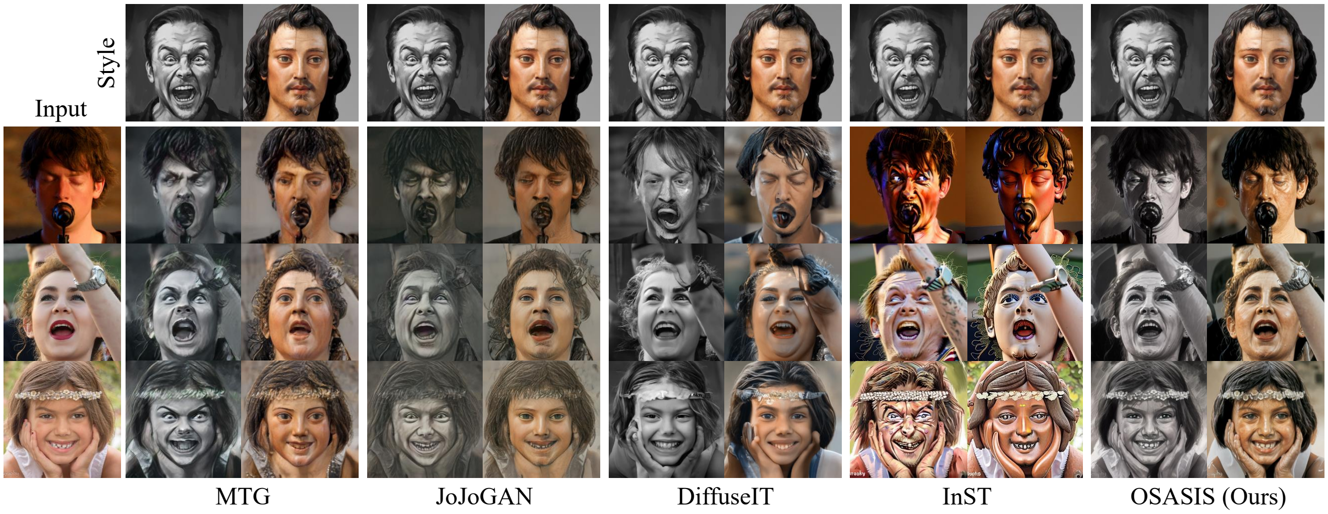


Figure 4. Comparison with other stylization methods. Note that our method successfully preserves the low-density attributes while other baseline methods fail to do so.

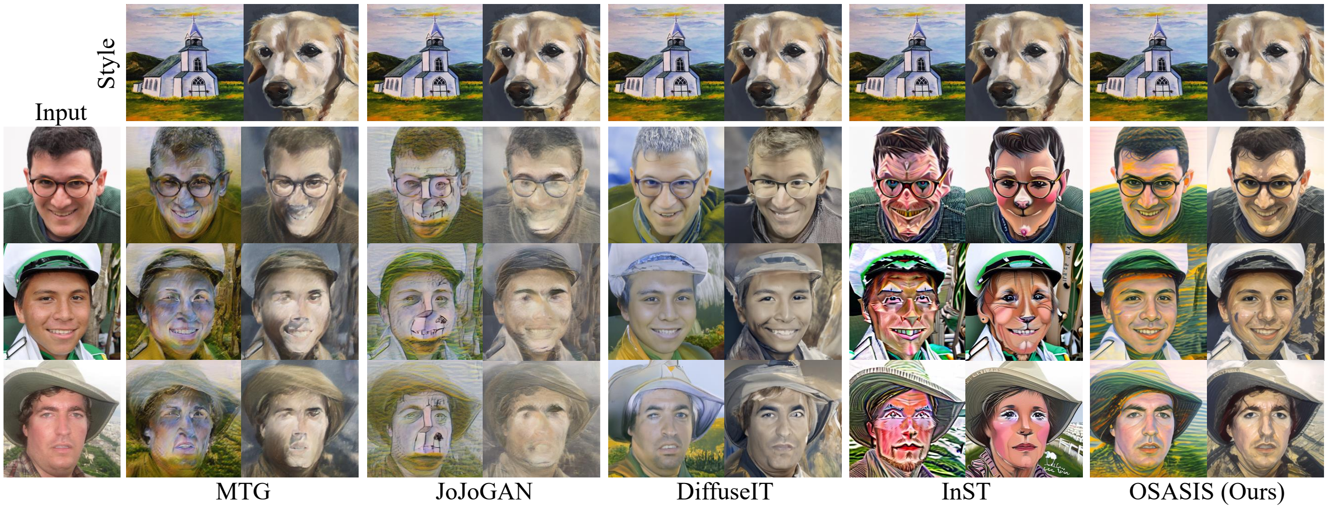


Figure 5. Stylization with OOD reference images. Due to the limited capabilities of GAN-based inversion methods, the baseline methods fail in disentangling the structure and semantics of the style image. This results in structural artifacts being transferred into the output image, whereas OSASIS successfully extracts only the semantics.

makes it difficult to transfer only the style factor from the reference image. As shown in Figure 5, OSASIS is able to stylize images with OOD reference images while preserving its content, whereas other baseline methods suffer from severe artifacts. Although DiffuseIT and InST manage to avoid severe artifacts, they still struggle with addressing domain gap and handling strong concept conditioning.

4.4. Stylization Results on Other Datasets

To evaluate the generalization capabilities of our method across various datasets, we executed stylization on several distinct datasets: AFHQ-dog [1], LSUN-church [34], and DeepFashion [19]. Figure 6 displays the efficacy and versatility of our approach across these diverse datasets. Notably, our method exhibits proficiency beyond facial stylization, adeptly adapting to various image domains.

	ArtFID↓	ID Sim↑	SD↓
w/o SPN	36.41	0.6595	0.0371
w SPN (λ_{SPN}=0.1)	**34.89**	0.6825	0.0361
w SPN (λ_{SPN}=0.5)	36.62	0.7177	**0.0348**
w SPN (λ_{SPN}=1.0)	43.15	**0.7321**	0.0355

Table 2. Quantitative ablation study of SPN

4.5. Stylization with Text-driven Manipulation

For text-driven manipulation, we directly optimize $\mathbf{z}_{\text{sem}}^{\text{in}}$ using CLIP directional loss. Once $\mathbf{z}_{\text{sem}}^{\text{in}}$ is optimized, it can be used to stylize the input image with the aforementioned mixing process. In Figure 7, we show our qualitative results of stylization with text-driven manipulation, where our model successfully incorporates the style of a reference image with manipulated attributes while being robust in content preservation.

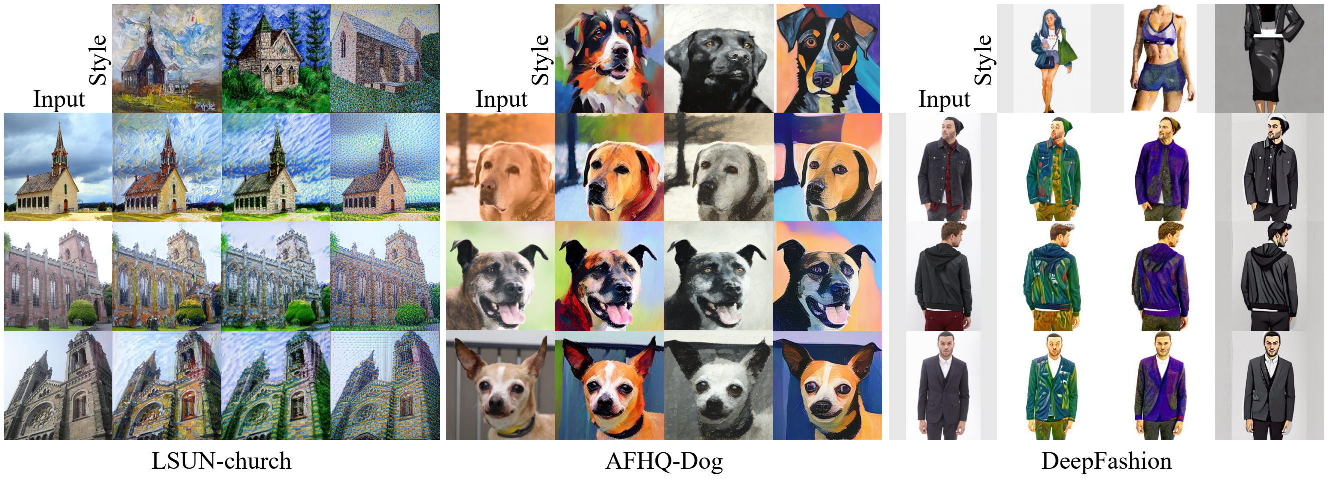


Figure 6. Stylization result of OSASIS on LSUN-church, AFHQ-dog, and DeepFashion.

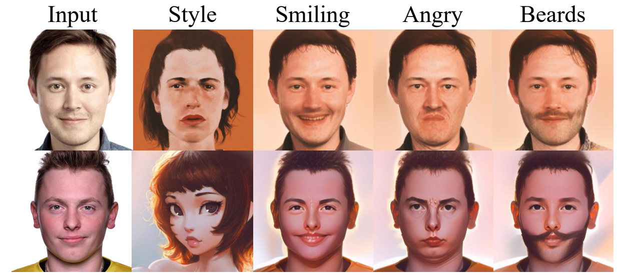


Figure 7. Stylization with text-driven manipulation. The optimized semantic code doesn't overwrite or harm the structure and style of the image, thus preserving the overall structure while manipulating attributes.

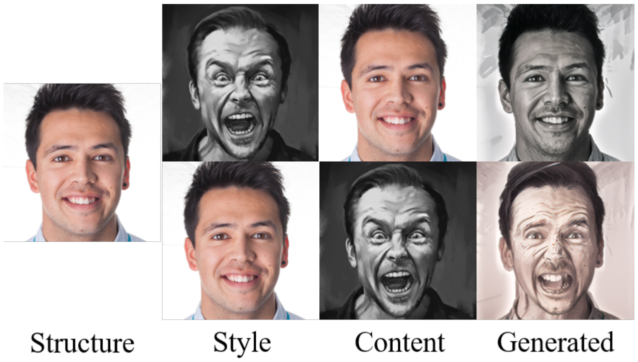


Figure 8. Ablation study of latent codes. The result shows that OSASIS is capable of effectively disentangling the structure from the semantics. By conditioning the semantic codes appropriately, we are able to control the content and style factors in the generated image.

4.6. Ablation Study

Latent Code Furthermore, we conduct ablation studies to shed light on the nature of mixing content and style into the feature maps of the UNet. In the first row of Figure 8, we perform normal stylization on an input image, *i.e.* by encoding its structural latent code $\mathbf{x}^{\text{in}}_{\mathbf{t}_0}$, conditioning $\mathbf{z}^{\text{style}}_{\text{sem}}$ to low-level feature maps, and conditioning $\mathbf{z}^{\text{in}}_{\text{sem}}$ to high-level feature maps. The second row shows the results of the semantic codes being conditioned oppositely. The resulting generated image 1) holds the structural integrity encoded from the structural latent code $\mathbf{x}^{\text{in}}_{\mathbf{t}_0}$, 2) preserves the content from the given content image (*i.e.* identity, facial expressions), and 3) retains the stylized attributes of the given style image (*i.e.* skin complexion). We show that by conditioning the semantic codes to their respective feature maps, OSASIS achieves control over mixing content and style.

Structure-Preserving Network Our SPN is implemented to aid the structural latent code $\mathbf{x}^{\text{in}}_{\mathbf{t}_0}$ in preserving the overall structure of the stylized image, due to the encoding process Eq.11 being formulated to add noise. Figure 9 shows the effects of our SPN, where the third column is a stylized sample without, and the fourth column is a stylized sample with SPN. It can be seen that the stochastic latent code $\mathbf{x}^{\text{in}}_{\mathbf{t}_0}$ faithfully preserves the overall structure of the image, but without SPN, objects along the edges of content images (e.g. hands, poles, fingers) are distorted. To investigate the effect of SPN further, we conduct a quantitative ablation study. Table 2 validates the efficacy of SPN, indicating substantial improvements in ID similarity and structure distance, thereby cementing its importance for maintaining content and structural integrity. It's important to note, however, that a λ_{SPN} value above 0.1 might overemphasize structural aspects, which can compromise the stylization quality, suggesting the necessity for careful calibration of λ_{SPN} to achieve an optimal balance in stylization.

5. Related Work

5.1. One-shot Image Stylization

Stylizing input images with only one reference image originates from neural style transfer, first introduced by Gatys *et al.* [6]. However, this method requires the stylized image to be optimized every generation, which was addressed

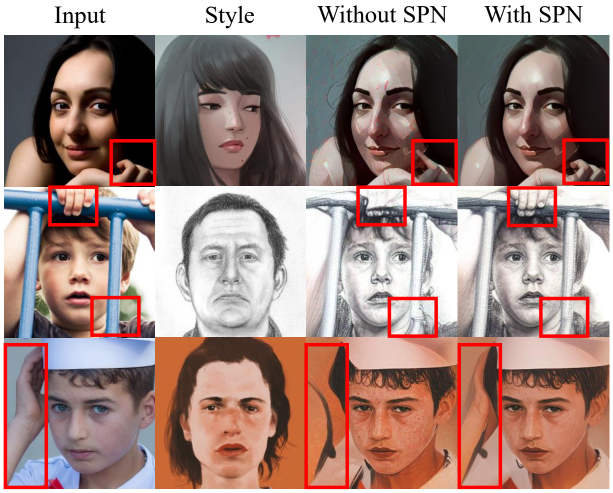


Figure 9. Ablation study of SPN. The results demonstrate that SPN is a crucial to ensure the preservation of the underlying structure while applying the stylization process.

by Johnson *et al*. [9] by introducing an image transform network for fast stylization. However, traditional NST methods are limited in their ability to capture the semantic information of input and style images.

For GAN-based models, one-shot image stylization is achieved by using one-shot adaptation methods, which aims to transfer a generator to a new domain using a single reference image. One-shot adaptation methods typically involve finetuning a generator using only a single reference image. Once the generator is finetuned, these methods can unconditionally generate stylized images, and by using GAN inversion techniques, they also achieve input image stylization. StyleGAN-NADA [5], while applicable to one-shot image stylization tasks, exhibits limited capability, primarily due to its initial development for text-driven style transfer purposes. The first successful GAN-based one-shot adaptation method is MTG [39], which uses CLIP directional loss to finetune the generator and mitigate overfitting problems. Other works have since focused on improving generation quality [16], content preservation [36], and entity transfer [38]. In contrast to one-shot adaptation approaches, our method aims to stylize real images with a single reference image instead of generating stylized synthesized images. Therefore, we refer to our method as one-shot image stylization. From our perspective, the work that is most comparable to our own is JoJoGAN [2]. JoJoGAN generates a training dataset by random style mixing and finetunes a generator to create a style mapper.

5.2. Image Manipulation with Diffusion Models

Image manipulation has advanced significantly in recent years, with methods presented by StyleGAN2 [12] being widely explored. However, the potential of diffusion models for high-quality image manipulation has been elucidated in recent research. DiffAE [23] introduces a semantic encoder that generates semantically meaningful latent vectors for diffusion models, which can be manipulated for attribute editing. DiffusionCLIP [13] demonstrates the effectiveness of diffusion-based text-guided image manipulation by finetuning a DDIM with CLIP directional loss. Asyrp [17] uncovers a semantic latent space in the architecture of diffusion models.The authors train the *h-space* manipulation module with CLIP directional loss, achieving consistent image editing results. While DiffusionCLIP and Asyrp employ CLIP directional loss, their focus is on text-guided manipulation whereas our work targets image-guided manipulation. DiffuseIT [15] aims to perform image translation guided by either text or image using a CLIP and a pretrained ViT [4]. Their approach leverages the reverse process of DDPM and incorporates CLIP and ViT to guide the image generation process. InST [37] employs textual inversion to extract the concept from a style image. By conditioning the generation process on this extracted concept, InST is able to stylize input images.

6. Conclusion

We have introduced OSASIS, a novel one-shot image stylization method based on diffusion models. In contrast to GAN-based and other diffusion-based stylization methods, OSASIS shows robust structure awareness in stylization, effectively disentangling the structure and semantics from an image. While OSASIS demonstrates significant advancements in structure-aware stylization, several limitations exist. A notable constraint of OSASIS is its training time, which is longer than comparison methods. This extended training duration is a trade-off for the method's enhanced ability to maintain structural integrity and adapt to various styles. Additionally, OSASIS requires training for each style image. This requirement can be seen as a limitation in scenarios where rapid deployment across multiple styles is needed. Despite these challenges, the robustness of OSASIS in preserving the structural integrity of the input images, its effectiveness in out-of-domain reference stylization, and its adaptability in text-driven manipulation make it a promising approach in the field of stylized image synthesis. Future work will address these limitations, particularly in optimizing training efficiency and reducing the necessity for individual style image training, to enhance the practicality and applicability of OSASIS in diverse real-world scenarios.

Acknowledgment We thank the ImageVision team of NAVER Cloud for their thoughtful advice and discussions. Training and experiments were done on the Naver Smart Machine Learning (NSML) platform [14]. This study was supported by BK21 FOUR.

References

[1] 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. 6

[2] Min Jin Chong and David Forsyth. Jojogan: One shot face stylization. In *European Conference on Computer Vision*, pages 128–152. Springer, 2022. 1, 5, 8

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 5

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8

[5] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 8

[6] 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. 7

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 5

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[9] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 8

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 4

[11] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 5

[12] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 8

[13] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 2, 3, 4, 8

[14] Hanjoo Kim, Minkyu Kim, Dongjoo Seo, Jinwoong Kim, Heungseok Park, Soeun Park, Hyunwoo Jo, KyungHyun Kim, Youngil Yang, Youngkwan Kim, et al. Nsml: Meet the mlaas platform with a real-world case study. *arXiv preprint arXiv:1810.09957*, 2018. 8

[15] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022. 2, 5, 8

[16] Gihyun Kwon and Jong Chul Ye. One-shot adaptation of gan in just one clip. *arXiv preprint arXiv:2203.09301*, 2022. 1, 8

[17] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 2, 8

[18] Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, and Wen Zheng. Blendgan: Implicitly gan blending for arbitrary stylized face generation. *Advances in Neural Information Processing Systems*, 34:29710–29722, 2021. 5

[19] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[20] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 3

[21] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021. 1

[22] Justin NM Pinkney and Doron Adler. Resolution dependent gan interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334*, 2020. 1

[23] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022. 2, 4, 8

[24] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1):1–13, 2022. 5

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[26] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2

[27] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[28] Guoxian Song, Linjie Luo, Jing Liu, Wan-Chun Ma, Chunpong Lai, Chuanxia Zheng, and Tat-Jen Cham. Agilegan: stylizing portraits by inversion-consistent transfer learning. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 1

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 2

[30] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 5

[31] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 5

[32] Matthias Wright and Björn Ommer. Artfid: Quantitative evaluation of neural style transfer. In *Pattern Recognition: 44th DAGM German Conference, DAGM GCPR 2022, Konstanz, Germany, September 27–30, 2022, Proceedings*, pages 560–576. Springer, 2022. 5

[33] Yiran Xu, Zhixin Shu, Cameron Smith, Seoung Wug Oh, and Jia-Bin Huang. In-n-out: Faithful 3d gan inversion with volumetric decomposition for face editing. *arXiv preprint arXiv:2302.04871*, 2023. 5

[34] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 6

[35] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4

[36] Yabo Zhang, Yuxiang Wei, Zhilong Ji, Jinfeng Bai, Wangmeng Zuo, et al. Towards diverse and faithful one-shot adaption of generative adversarial networks. In *Advances in Neural Information Processing Systems*, 2022. 1, 8

[37] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10156, 2023. 2, 5, 8

[38] Zicheng Zhang, Yinglu Liu, Congying Han, Tiande Guo, Ting Yao, and Tao Mei. Generalized one-shot domain adaption of generative adversarial networks. *arXiv preprint arXiv:2209.03665*, 2022. 1, 8

[39] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. In *International Conference on Learning Representations*, 2021. 1, 2, 4, 5, 8