This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Balanced Image Stylization with Style Matching Score

Yuxin Jiang[1,2] Liming Jiang[3] Shuai Yang[4] Jia-Wei Liu[1] Ivor W. Tsang[2,3] Mike Zheng Shou[1†]

[1]Show Lab, National University of Singapore
[2]CFAR & IHPC, Agency for Science, Technology and Research (A*STAR)
[3]Nanyang Technological University [4]Wangxuan Institute of Computer Technology, Peking University

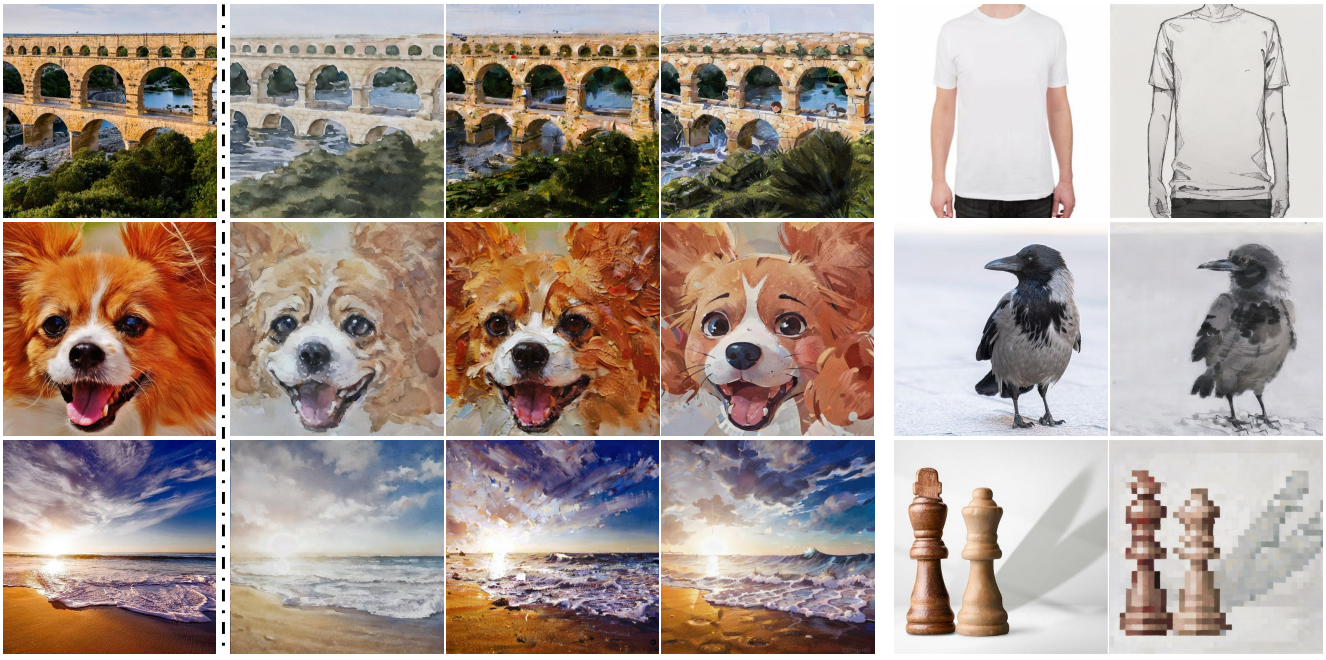


Figure 1. **Examples of SMS applied to various artistic styles.** Stylized outputs (left to right, top to bottom) include watercolor, oil painting, kids' illustration, sketch, chinese ink painting, and pixel art style.

Abstract

We present Style Matching Score (SMS), a novel optimization method for image stylization with diffusion models. Balancing effective style transfer with content preservation is a long-standing challenge. Unlike existing efforts, our method reframes image stylization as a style distribution matching problem. The target style distribution is estimated from off-the-shelf style-dependent LoRAs via carefully designed score functions. To preserve content information adaptively, we propose Progressive Spectrum Regularization, which operates in the frequency domain to guide stylization progressively from low-frequency layouts to high-frequency details. In addition, we devise a Semantic-Aware Gradient Refinement technique that leverages relevance maps derived from diffusion semantic priors to selectively stylize semantically important regions. The proposed optimization formulation extends stylization from pixel space to parameter space, readily applicable to lightweight feed-forward generators for efficient one-step stylization. SMS effectively balances style alignment and content preservation, outperforming state-of-the-art approaches, verified by extensive experiments. Code: https://github.com/showlab/SMS.

1. Introduction

Image stylization transforms a content image to adopt a specific visual style, with applications in art design [5], virtual reality [9], and content creation [4, 23, 24, 27]. Early exemplar-based methods like Neural Style Transfer

† Corresponding author.

(NST) [5, 13] capture style using statistical correlations in CNN feature maps of a style image but are limited to single-image styles and are heavily content-influenced. Collection-based methods using GANs [2, 37] model style as distributions over datasets for more realistic transfer but require large, well-curated datasets and face issues like mode collapse and training instability.

The emergence of multimodal models like CLIP [21], and diffusion models (DMs), such as Stable Diffusion (SD) [22], has ushered in a new era for image style transfer. Recent approaches can be categorized into zero-shot text-guided, one-shot exemplar-based, and collection-based fine-tuning methods. Zero-shot text-guided methods [6, 32] use text prompts for style transfer but often struggle to accurately capture complex styles that are challenging to describe with words (see Figure 2(a)), as "one image is worth a thousand words". One-shot exemplar-based methods [3, 28, 36], similar to traditional NST, struggle to model an artist's overall style and often overemphasize the exemplar's style, leading to content distortion (see Figure 2(b)). Collection-based fine-tuning methods, such as those using LoRA [10], adapt DMs to new styles with minimal retraining, effectively capturing intricate patterns and an artist's overall style. However, these methods often overemphasize style at the expense of content preservation, struggling to balance style and content. When conditioned on specific content images, they typically use ControlNet [34] with edge maps as conditions or DDIM inversion [26] and generate stylized images from noises, which may still fail to preserve sufficient content details (see Figure 2(d)).

To address these challenges, we propose the *Style Matching Score (SMS)*, a novel optimization method that reframes style transfer as style distribution matching. Specifically, we introduce: *1) Style Matching Objective* for transferring style; *2) Progressive Spectrum Regularization* for preserving content; and *3) Semantic-Aware Gradient Refinement* for balancing style and content with diffusion semantic priors. Our key idea is to match the distribution of the output images with the target style distribution by leveraging powerful diffusion priors. We minimize the Kullback–Leibler (KL) divergence between the distribution of the stylized images and the target style distribution, estimated by score functions of a style LoRA-integrated pretrained DM. By interpreting the denoised gradient directions that make the image more stylized, we effectively align generated images with the target style. At a high level, our method shares the motivation of score distillation methods [29, 33]. We differ by focusing on distilling style information into the real image domain, where preserving the identity of the source content is a key challenge.

To achieve this, we incorporate an explicit progressive identity regularization in frequency domain. Observing that stylistic differences largely reside in high-frequency components, we guide stylization progressively from low-level structures to high-frequency details, ensuring structural integrity and harmonious transitions. This is achieved by preserving more low-frequency components during high-noise steps and allowing finer stylized details at lower-noise steps. Combined with an adaptive narrowing sampling strategy, our method minimizes content disruption while retaining sufficient style details. To further balance style and content, we introduce a semantic-aware gradient refinement that leverages the relevance map [18] derived from DM's semantic priors. Our key insight is that each pixel requires different degrees of stylization. By uniquely using the map as an element-wise weighting mechanism, we allow gradients in more semantically-important regions to emphasize stylistic transformation while attenuating gradients in less critical areas to preserve content integrity.

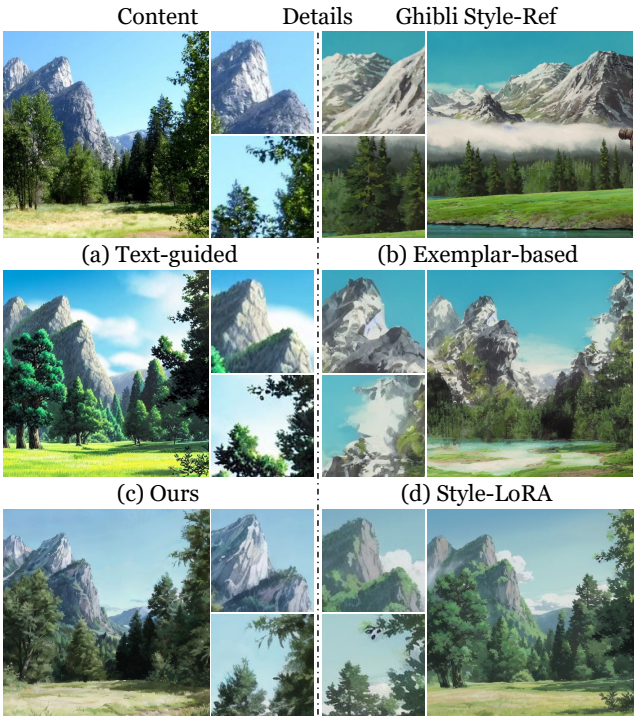


Figure 2. **Comparison of different style representations.** (a) Fails to mimic the Ghibli style. (b) Introduces style elements but lack spatial coherence, simply overlaying textures from the reference style. (d) Captures Ghibli-like textures but distorts content and color due to limitations in ControlNet [34] conditioning. (c) Achieves a harmonious balance of content preservation and style adaptation with fine local stylized details.

Thanks to our optimization-based formulation, SMS extends stylization from the pixel space to the parameter space, enabling broader and more flexible style transfer capabilities. For instance, SMS is readily applicable to a lightweight feed-forward generator, allowing for efficient one-step generation, potentially useful for more applications beyond. In summary, our contributions are fourfold:

- We introduce a novel SMS optimization method to leverage diffusion priors to match style distributions for semantic-aware and content-preserving style transfer.
- We develop a progressive spectrum regularization term and adaptive narrowing sampling strategy guided by diffusion characteristics, to preserve content and achieve fine-detail stylization.
- We propose an semantic-aware gradient refinement to balance style and content by selectively stylizing semantically-important regions.
- We demonstrate SMS is readily applicable to distill style into lightweight feed-forward networks, extending applicability beyond single-image stylization.

2. Related Work

Image Style Transfer aims to synthesize images in artistic styles and has significant practical value. GAN-based methods [2, 12, 37] enable realistic stylization by learning style distributions from datasets, which however requires extensive data and are challenging to train. Recent work leverages pretrained diffusion models (DMs) for style transfer, categorized as follows. *1) Prompts-Guide Stylization*, represented by FreeStyle [6], offers flexibility but struggles with ambiguous or imprecise style control due to the abstract nature of text guidance. *2) Exemplar-Based Stylization* typically manipulates self-attention or cross-attention layers to transfer style from a reference image. InST [36] maps style images to textual embeddings for conditions. StyleID [3] maintains content integrity through query preservation and AdaIN [11]. InstantStyle-Plus [28] injects style-specific features into intermediate model layers. Single-style exemplars, however, lead to over-stylization, obscuring semantic content. *3) Tuning-Based Stylization* fine-tunes DMs on target style distributions, which produces high-quality results but is computationally costly. Style-dependent LoRA adapt models to specific styles with less data and capture nuanced style distributions [1]. However, they typically use ControlNet [34] with basic edge maps from the source image for image-to-image stylization, which may lose details. To overcome these limitations, we propose a distribution-level style matching loss, fully leveraging diffusion priors for semantic-aware, identity-preserving and applicable stylization.

Score Distillation, represented by Score Distillation Sampling (SDS) [20] and Variational Score Distillation (VSD) [29], is initially introduced in the context of text-to-3D synthesis by distilling the 2D generative prior of text-to-image DMs. Later, researchers have applied this idea to diffusion distillation, such as Distribution Matching Distillation (DMD) [33] and text-driven image editing [7, 16, 19]. However, these SDS-based image editing methods mainly overemphasize the style and fail to preserve the content for image stylization. To solve this issue, Delta Denoising Score (DDS) [7] reduces noisy gradient directions in SDS to better maintain the input image details. Building on DDS, Posterior Distillation Sampling (PDS) [16] introduces a stochastic latent matching loss to add an explicit spatial identity preservation regularization term, while contrastive Denoising Score (CDS) [19] adds a contrastive loss to maintain identity. Despite efforts, DDS suffers from content preservation due to increasing source-target distribution mismatch [17], as reduced noisy denoising directions are not calculated based on the current optimized image. In contrast, VSD effectively mitigates this mismatch by test-time finetuning a copy of the DM with LoRA on the current set (*i.e.*, the fake distribution). Our method shares the high-level objectives as VSD and DMD but prioritizes identity preservation for stylization. We introduce a semantic-aware gradient refinement and spectrum regularization, specifically tailored for style transfer, ensuring precise stylization.

3. Style Matching Score

3.1. Problem Modeling

Given a source image $x^{\text{src}} \in p_{\text{real}}$, a text description y_{src}, and a fixed target style distribution p_{style}, our goal is to synthesize a stylized image x^{tgt} such that:

1. *Style Alignment:* aligns with the style distribution p_{style}.
2. *Content Preservation:* retains the identity of x^{src}.

To model this problem flexibly, we parameterize the image generation process using θ. Specifically, we define a parametric generator G_θ such that $x^{\text{tgt}} = G_\theta(x^{\text{src}})$. Here, θ can represent either the image itself (in test-time optimization) or the parameters of an image-to-image generator network.

To achieve style alignment, we aim to minimize the KL divergence between the generated distribution p_{G_θ} and target style distribution p_{style}:

$$D_{\text{KL}}(p_{G_\theta} || p_{\text{style}}) = \int p_{G_\theta}(x^{\text{tgt}}) \log \frac{p_{G_\theta}(x^{\text{tgt}})}{p_{\text{style}}(x^{\text{tgt}})} dx. \quad (1)$$

3.2. Style Matching Objective

Following DMD [33], we estimate the distributions in Eq. (1) using score functions approximated by diffusion models. Specifically, we employ two noise prediction models ϵ_{style} (fixed) and $\epsilon^{\phi}_{\text{fake}}$ (dynamically learned with $L^{\phi}_{\text{denoise}}$ [33]), which approximate the score functions of p_{style} and p_{G_θ}, respectively. Then, $\nabla_\theta D_{\text{KL}}$, the gradient of D_{KL} with respect to θ, is approximated as:

$$\mathop{\mathbb{E}}_{t,\epsilon}\left[w_t \left(\epsilon_{\text{style}}(z^{\text{tgt}}_t; y_{\text{src}}, t) - \epsilon^{\phi}_{\text{fake}}(z^{\text{tgt}}_t; y_{\text{src}}, t) \right) \frac{\partial G_\theta}{\partial \theta} \right], \quad (2)$$

where y_{src} is the text prompt describing the content of the source image x^{src}, $z^{\text{tgt}}_t = \sqrt{\bar{\alpha}_t} z^{\text{tgt}}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, and $z^{\text{tgt}}_0 = \varepsilon(G_\theta(x^{\text{src}}))$. The function $\varepsilon(\cdot)$ is the VAE encoder of the SD, and w_t is a timestep-dependent scalar

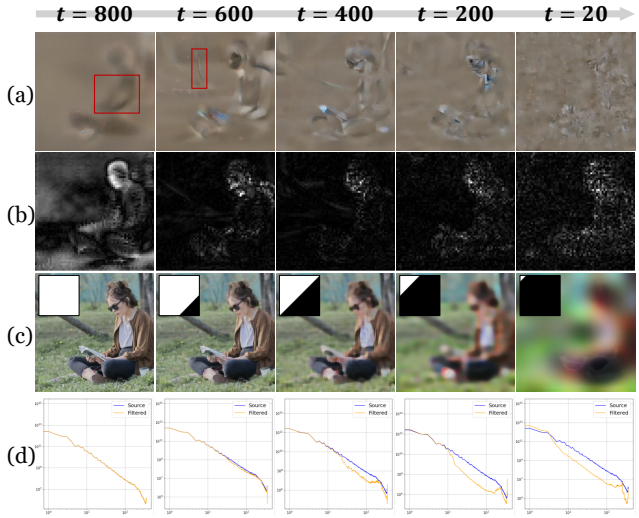


Figure 3. **Visualization across different timesteps.** (a) Style matching gradient $\nabla_\theta D_{\text{KL}}$; (b) Relevance Coefficient $\mathcal{R}(z_t^{\text{src}}, t)$; (c) Regulated frequency component in the spatial domain IDCT($\mathcal{F}_{\text{low}}$) with the corresponding low-pass filter mask LPF($\cdot$); (d) Radial average power spectrum comparing the source image's frequency distribution (blue) with the regulated component (orange).

weight. Intuitively, the denoising direction ϵ_{style} moves z_0^{tgt} toward the modes of p_{style}, while $-\epsilon_{\text{fake}}^{\phi}$ spreads them apart, steering the image towards a desired stylized direction. For a detailed derivation, please refer to [33] and our supplementary material.

Our approach differs from DMD [33], which employs a general pretrained DM ϵ_{real} without style specialization to represent the target distribution. In contrast, we tailor the DM to the target style by integrating a style-specific LoRA module into ϵ_{real}, resulting in ϵ_{style}. This integration harnesses the superior style modeling capabilities of the style-LoRA, bridging the gap between general DM and specific style needs. Our method captures the nuances of the target style more accurately, enhancing stylization fidelity and better aligning the generated images with the desired artistic characteristics.

3.3. Progressive Spectrum Regularization

Unlike DMD, which maps from noise to images $z_{\theta,0} = \varepsilon(G_\theta(\eta))$ with $\eta \sim \mathcal{N}(0, \mathbf{I})$, our style matching objective focuses on transforming the real image distribution to the style distribution, required preserving content correspondence. However, because the text prompt y_{src} allows for a probabilistic distribution of all possible outputs, the raw gradient $\nabla_\theta D_{KL}$ can be noisy, introducing unrelated changes (see Figure 3(a)).

To mitigate the loss of fidelity, we introduce explicit content regularization in frequency domain. The analysis (see our supplementary material) indicates that the primary differences between real and stylized images lie in the high-frequency components. Previous methods that directly regulate in the spatial domain either fail to preserve identity or apply too strict regularization, limiting effective style transfer. In contrast, our method progressively guide stylization from low-level structures to high-frequency details using a timestep-aware spectrum. Specifically, we preserve more low-frequency components during high-noise steps (large t) for structural integrity and allow greater flexibility in high-frequency components during low-noise steps (small t) to enable detailed stylization. The progressive spectrum regularization term is defined as:

$$L_{\text{freq}} = \Big|\Big|\mathcal{F}_{\text{low}}(z_0^{\text{tgt}}, t) - \mathcal{F}_{\text{low}}(z_0^{\text{src}}, t)\Big|\Big|_2^2, \tag{3}$$

where $\mathcal{F}_{\text{low}}(z, t) = \text{LPF}(\text{DCT}(z), \text{thld}(t))$, where LPF denotes a low-pass filter. Specifically, LPF$(\cdot, \text{thld}(t))$ applies a low-pass filter to the Discrete Cosine Transform (DCT) of z using a cutoff frequency threshold determined by t, denoted as thld(t). As shown in Figure 3(c), thld(t) decreases with t, resulting in progressively looser regularization for step-aware content preservation.

3.4. Semantic-Aware Gradient Refinement

To further balance style and content, we introduce a semantic-aware gradient refinement mechanism. Recognizing that not every pixel of an image require the same degree of stylization; for example, foreground subjects might need stronger stylization, whereas background elements benefit from minimal alteration to preserve identity, we leverage the DM's semantic priors to computer a relevance map [18] to guide gradient modulation. Specifically, we define the relevance coefficient $\mathcal{R}(z_t^{\text{src}}, t)$ at timestep t as:

$$\mathcal{R}(z_t^{\text{src}}, t) = \text{Norm}\Big(\big|\epsilon_{\text{real}}(z_t^{\text{src}}; y_{\text{edit}}, t) - \epsilon_{\text{real}}(z_t^{\text{src}}; y_\emptyset, t)\big|\Big), \tag{4}$$

where $z_t^{\text{src}} = \sqrt{\bar{\alpha}_t} z_0^{\text{src}} + \sqrt{1-\bar{\alpha}_t}\epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, and $z_0^{\text{src}} = \varepsilon(x^{\text{src}})$. y_{edit} is the editing instructions specifying the desired style, and $y_\emptyset$ is the empty condition. The absolute difference $|\cdot|$ between these predictions highlights regions where the model anticipates changes due to the style instructions. The function Norm$(\cdot)$ applies min-max normalization to scale the relevance map to range $[0, 1]$.

This relevance map effectively captures the semantic importance of each pixel with respect to the style transformation. By employing it as an element-wise weighting mechanism, we modulate the gradient to emphasize stylistic changes in semantically significant regions while attenuating changes in less critical areas. The refined style matching loss is:

$$L_{\text{style}} = \mathbb{E}_{t,\epsilon}\Big[\Big|\Big|\mathcal{R}(z_t^{\text{src}}, t) \odot w_t\Big(\epsilon_{\text{style}}(z_t^{\text{tgt}}; y_{\text{src}}, t) - \epsilon_{\text{fake}}^{\phi}(z_t^{\text{tgt}}; y_{\text{src}}, t)\Big)\Big|\Big|_2^2\Big], \tag{5}$$

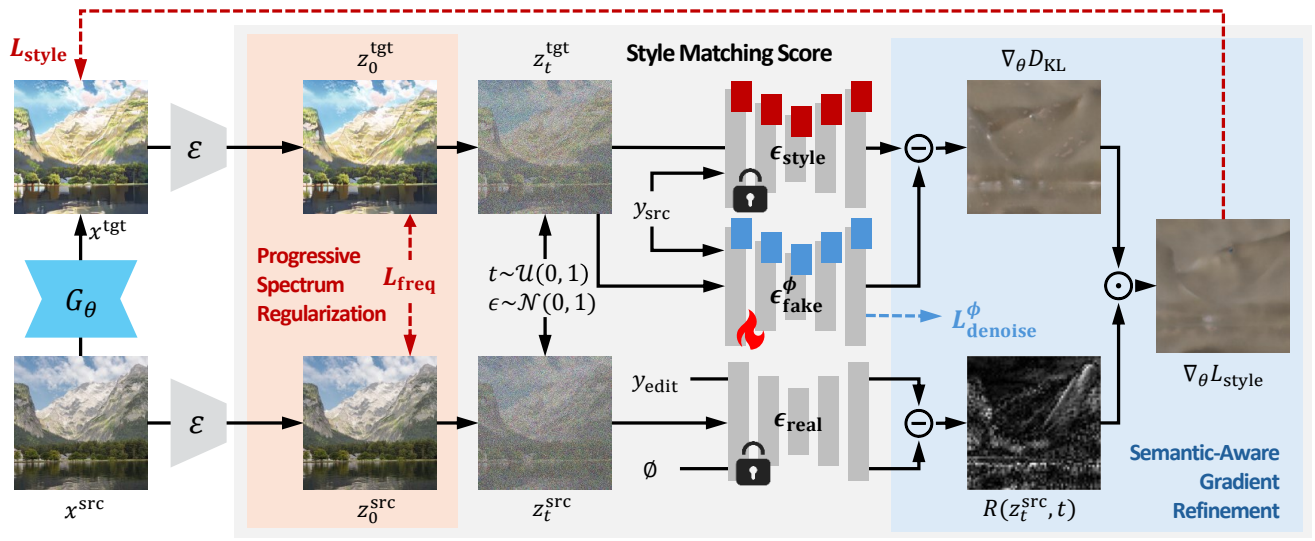


Figure 4. **Overview of SMS for Feed-Forward Stylization.** We train a one-step generator G_θ to map source images x^{src} into stylized outputs. To transfer style, we compute the style matching gradient $\nabla_\theta D_{\text{KL}}$ by injecting random noise into the generated image and passing it through two diffusion models: one integrated with Style-LoRA (representing the target style distribution) and one continually trained on the generated images (representing the current "fake" distribution). The difference between their denoising scores provides a gradient direction to make the image more stylized and less fakeness. A relevance map, derived from the disagreement between noise predictions with and without stylized instructions, acts as an element-wise correction to selectively stylize semantically important regions, ensuring fidelity and coherence. Finally, progressive spectrum regularization is applied to preserve content within the latent space.

where $\odot$ denotes element-wise multiplication. Moreover, the relevance coefficient is adaptive, timestep-depedent, it aligns with the progression of the diffusion process, focusing gradient modulation on appropriate regions at different step (see Figure 3(b)). This dynamic adjustment allows for controlled and semantic-aware coherent style transfer across optimization.

3.5. Overall Training Objective

Combining the progressive spectrum regularization and semantic-aware gradient refinement, our overall training objective is:

$$L_{\text{SMS}} = L_{\text{style}} + \lambda \cdot L_{\text{freq}}, \tag{6}$$

where λ is the loss weight. By minimizing L_{SMS}, we guide G_θ to produce images that align with the target style in semantically important regions while retaining the identity of the source image.

Adaptive Narrowing Sampling Strategy. Uniformly random sampling of timesteps t results in inconsistent regularization strength across iterations, leading to blurred images and ineffective stylization. Conversely, naive linear annealing of t cause excessive high-frequency editing in later stages, producing deviations from the original content due to looser content restrictions. To address this, we propose an adaptive narrowing sampling strategy that gradually reduces the upper bound of the timestep sampling range over iterations while preserving randomness, effectively integrating with our progressive spectrum regularization. Specifically, we sample t as:

$$t \sim \mathcal{U}(t_{\min}, t_{\text{upper}}), \quad t_{\text{upper}} = \left(1 - \frac{\text{iter}_{\text{cur}}}{\text{iter}_{\text{total}}}\right) \cdot t_{\max}, \tag{7}$$

where iter_{cur} is the current iteration and $\text{iter}_{\text{total}}$ is the total number of iterations.

4. Feed-Forward Stylization with SMS

As an applications of SMS, we present a detailed data-free stylizaton pipeline using a lightweight feed-forward network G_θ (see Figure 4). Given a set of real images $\{x_i\}_{i=1}^N \in p_{\text{real}}$ and a target style domain p_{style} modeled by a pretrained ϵ_{style}, our goal is to learn a mapping $G_\theta : p_{\text{text}} \rightarrow p_{\text{style}}$ that transforms input images into the target style domain.

Reconstruction Warmup. To ensure that G_θ maintains content fidelity and accelerates convergence, we initialize G_θ by training it to reconstruct the input images. This warmup phase aligns the initial generated distribution p_{G_θ} with the real image distribution p_{real}, providing a solid foundation for subsequent training. The reconstruction loss is defined as: The reconstruction loss is defined as:

$$\mathcal{L}_{\text{rec}} = \mathop{\mathbb{E}}_{x \sim p_{\text{real}}} \left[\left|\left| G_\theta(x) - x \right|\right|_1 \right], \tag{8}$$

where $||\cdot||_1$ denotes the $L1$ norm.

Per-Batch Variable Timesteps. After warmup, we train G_θ using the SMS loss (see Eq. (6)). Unlike direct pixel updates in single-image optimization, updates to G_θ affect the

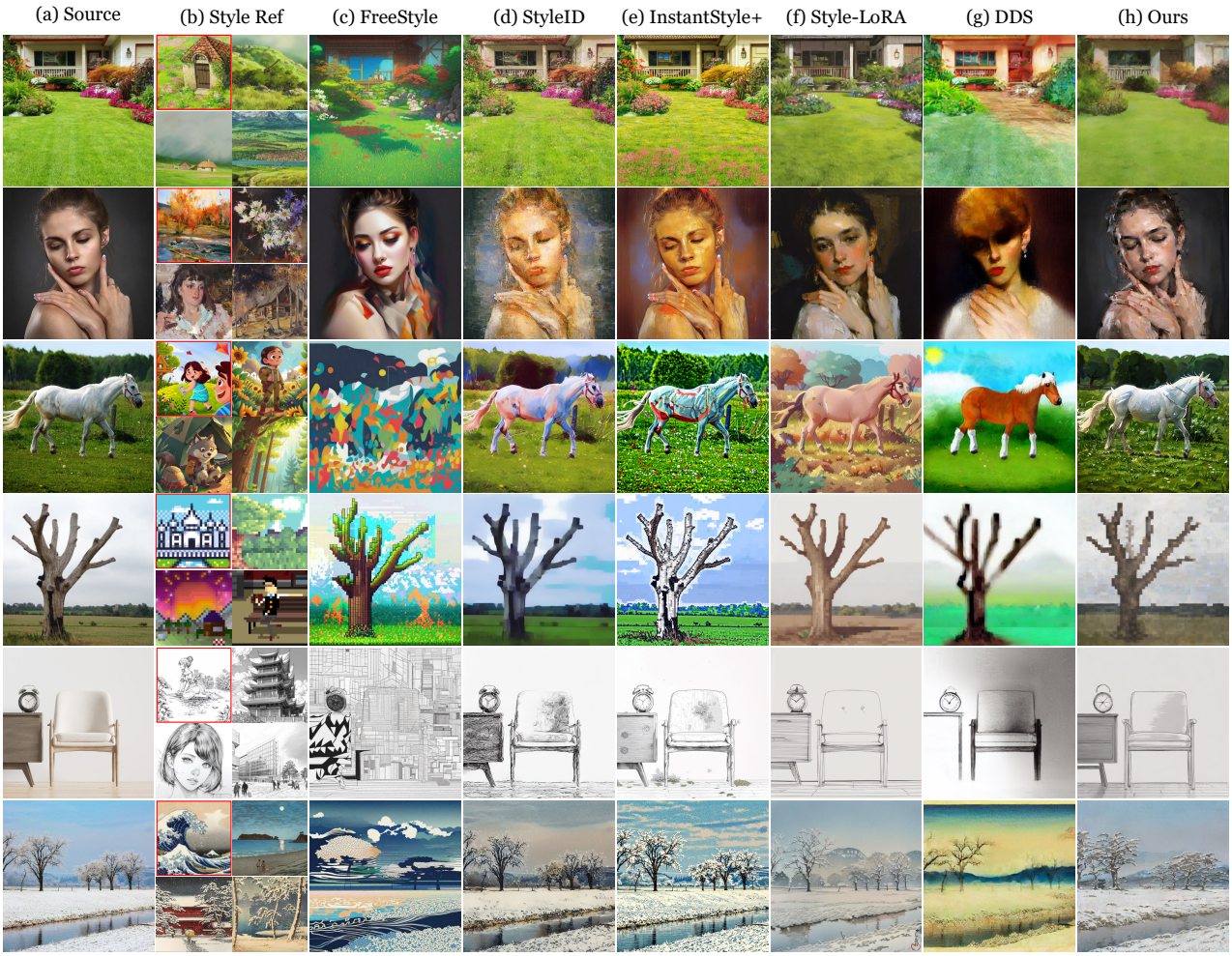


Figure 5. **Qualitative comparison.** We compare SMS (Ours) with five representative methods. Our approach achieves superior semantic consistency and style texture fidelity, striking the best balance between style alignment and content preservation compared to state-of-the-art baselines. The style references in the red boxes are used as the exemplars for exemplar-guided methods. Please zoom in for details

Table 1. Quantitative comparison with diffusion-based baselines using different style representations. The best results are highlighted in **bold**, and the second-best are underlined.

Metric	Ghibli Style							Oil Painting Style						
	Real	FreeStyle [6]	StyleID [3]	InstantStyle -Plus [28]	Style -LoRA	DDS [7]	Ours	Real	FreeStyle [6]	StyleID [3]	InstantStyle -Plus [28]	Style -LoRA	DDS [7]	Ours
LPIPS ↓	0.000	0.690	0.608	0.538	0.438	0.513	**0.326**	0.000	0.686	0.626	0.524	0.588	0.569	**0.431**
CFSD ↓	0.000	0.786	0.139	0.337	0.113	0.129	**0.102**	0.000	1.083	0.156	0.138	0.129	0.130	**0.128**
FID ↓	15.138	12.361	19.007	14.949	**12.267**	15.233	13.089	30.124	29.974	**16.252**	24.858	21.861	23.133	17.905
ArtFID ↓	16.138	22.582	32.169	24.532	19.077	24.554	**18.686**	31.124	52.218	28.046	39.402	36.292	37.864	**27.055**
PickScore ↑	1.000	0.683	0.405	1.019	**2.067**	0.537	1.487	1.000	0.247	0.803	1.350	**2.862**	0.782	1.842
HPSv2 ↑	0.272	0.201	0.210	0.257	**0.264**	0.227	0.259	0.260	0.164	0.224	0.258	**0.278**	0.231	0.260

generated distribution indirectly through changes in the network parameters. Additionally, network training requires generalization to diverse inputs and robustness to variations. To address these, we introduce per-batch variable timesteps. For each training batch, we replicate the same image x B times (where B is the batch size). We then add noise at different timesteps $\{t_i\}_{i=1}^{B}$ to each instance in the batch:

$$z_{t_i} = \sqrt{\bar{\alpha}t_i} z_0 + \sqrt{1-\bar{\alpha}t_i}\epsilon_i, \tag{9}$$

where $z_0 = \mathcal{E}(G_\theta(x))$. $\epsilon_i \sim \mathcal{N}(0, \mathcal{I})$ are independent noise samples, and t_i are independently sampled timesteps. By computing gradients over multiple noise levels for the same image, we effectively average the optimization directions, leading to more stable and consistent updates to G_θ.

5. Experiments

5.1. Single-Image Stylization

Baselines. We compare our method against five state-of-the-art diffusion-based style transfer methods categorized

by their style representations: 1) Zero-shot text-driven: FreeStyle [6], DDS [7]; 2) One-shot Exemplar-guided: StyleID [3], InstantStyle-Plus [28]; 3) Collection-based: Style-specific LoRA generation with ControlNet integrated for spatial preservation (referred to as Style-LoRA).

Datasets. We use the PIE-benchmark [14], containing 280 diverse images covering animals, people, indoor and outdoor scenes, each paired with textual descriptions.

Implementation Details. We conduct all experiments using Stable Diffusion 1.5. The relevance map's editing instruction y_{edit} is set as "Turn it into {*target style*}", with a total of 500 training iterations. The DCT threshold frequency thld(t) is defined as $t/500$. Sampling t ranges from $[20, 500]$ following our adaptive narrowing sampling strategy. Baseline methods are implemented using their official code and default settings. More implementation details are provided in the supplementary material.

Evaluation Metrics. We quantitatively evaluate our method on two representative artistic styles: *"Ghibli style"*, a renowned Japanese anime style known for its intricate semantic-aware style textures, and *"Oil Painting style"*, characterized by its coarser, larger brush strokes. We assess style transfer performance across three aspects: content preservation, style fidelity, and overall image quality. Content preservation is measured using LPIPS [35] and CFSD [3], evaluating structural similarity between stylized and corresponding source images. Style fidelity is quantified via FID [8] against authentic Ghibli images from original films. Overall perceptual quality is evaluated by ArtFID [30], defined as $(\text{LPIPS} + 1) \cdot (\text{FID} + 1)$, alongside aesthetic metrics such as PickScore [15] and Human Preference Score (HPS) [31]. Additionally, we conduct a user study to further validate the balance and effectiveness of our method. Participants are asked to choose the best result in terms of three criteria: evident style (Style), content preservation (Content), and overall translation quality (Balance). Higher scores indicate better image quality.

Qualitative Comparison. As illustrated in Figure 5, our approach excels at balancing style and content. FreeStyle over-shifts the style, losing identity preservation. Exampler-based methods maintain content integrity through DDIM inversion but lack sufficient style infusion, appearing as texture overlays and noticable color shifts. Style-LoRA effectively captures stylistic textures but struggle with content and color preservation due to limitations in ControlNet conditioning. DDS faces challenges in transferring style using text prompts alone, leading to content deviations and oversaturated colors due to the absence of explicit identity regularization. SMS, in contrast, presents delicate, semantic-aware style features while retaining strong identity preservation. The qualitative results indicate the effectiveness of our method, surpassing the state-of-the-art baselines in perceptual quality for single-image stylization.

Table 2. **User preference scores.**

Metric	FreeStyle	StyleID	InstantStyle+	Style-LoRA	DDS	Ours
Style	0.060	<u>0.147</u>	0.083	0.100	0.033	**0.577**
Content	0.003	0.127	<u>0.136</u>	0.090	0.017	**0.627**
Balance	0.013	0.110	<u>0.127</u>	0.077	0.020	**0.653**

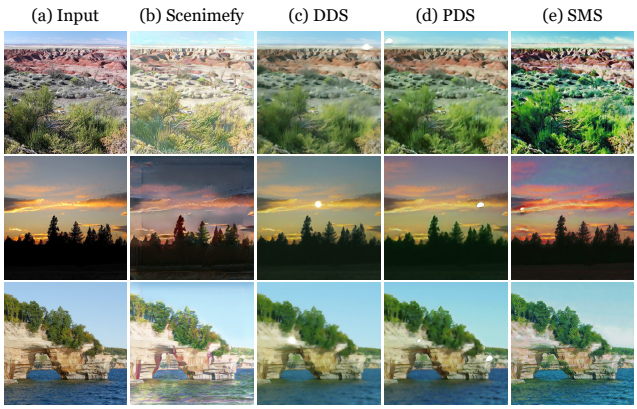


Figure 6. **Feed-Forward stylization comparison.**

Table 3. **Feed-Forward quantitative comparison.** The best results are in **bold**, and the second-best are <u>underlined</u>.

Metric	Real	Scenimefy [12]	DDS [7]	PDS [16]	SMS
LPIPS ↓	0.000	0.422	<u>0.321</u>	0.427	**0.268**
FID ↓	12.608	**12.054**	12.886	16.229	<u>12.472</u>
ArtFID ↓	13.608	18.561	<u>18.338</u>	24.590	**17.079**

Quantitative Results. Table 1 presents a quantitative evaluation of our method against baselines. For reference, we include metrics computed between the content and respective style datasets, labeled as "Real". Our method surpasses baselines in context preservation, with significantly lower LPIPS and CFDS. Regarding style representation, SMS attains competitive FID scores and secures second-best results in PickScore and HPS, indicating strong visual appeal. While Style-LoRA performs slightly better in certain style metrics, this is expected, as our target style distribution aligns with Style-LoRA's generation distribution, serving as the upper bound for our method. Despite this, SMS effectively balances content and style, achieving the best ArtFID score across intricate and coarse artistic styles, further showcasing its robustness and versatility.

User Preference Score. A total of 300 comparisons from 20 participants over 15 cases across 6 styles are collected. Table 2 summarizes the average preference scores, where we achieve the highest ratings in all three criteria, fruther suggesting the effectiveness of our method. Moreover, the results confirm that users prefer outcomes with better alignment to original content.

5.2. Feed-Forward Stylization

Experimental Settings. We train a lightweight feed-forward network G_θ (∼43MB) using SMS for real-time inference and compare it with state-of-the-art GAN-based anime stylization methods like Scenimefy [12],

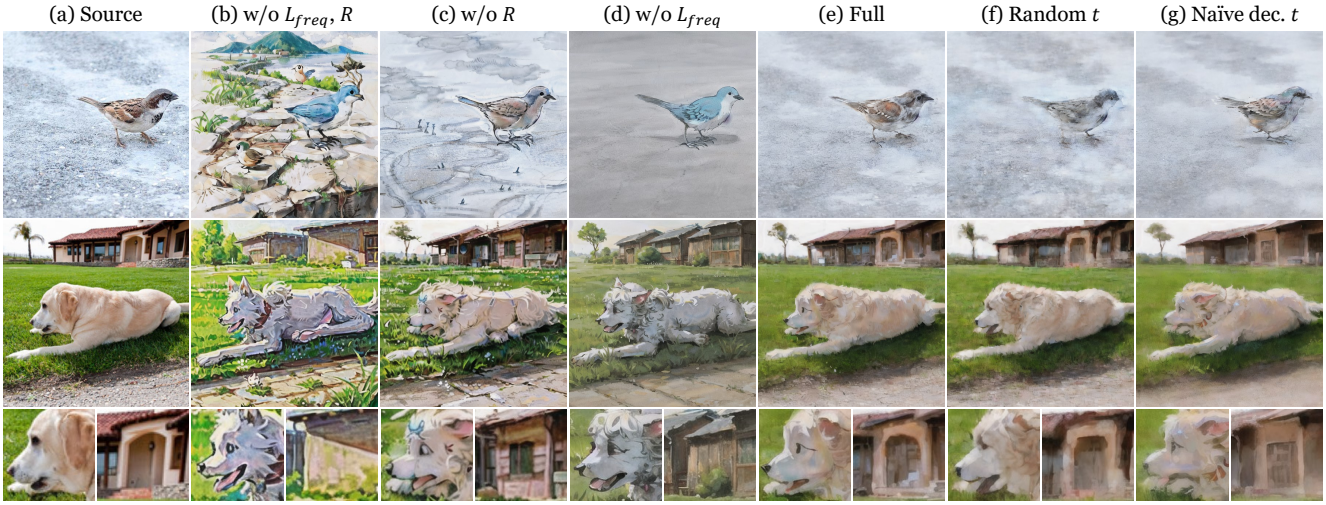


Figure 7. **Ablation studies of SMS.** The effect of each key component is illustrated.

and diffusion-based score optimization methods such as DDS [7] and PDS [16]. Our SMS setup follows the single-image experiments in Section 5.1, using a batch size B of 4. We train on 60k LHQ [25] natural scenes at 512×512 resolution, with a test set of $1,000$ sampled images from the remaining dataset.

Qualitative Comparison. Figure 6 shows that while Scenimefy produces visually pleasing results by leveraging real style data, it suffers from artifacts due to the instability inherent in GAN training. DDS and PDS, reliant on text-driven prompts, fail to achieve effective style transfer, leading to blurry outputs and weak style alignment. In contrast, SMS successfully distills styles into the model, capturing detailed stylistic features with fewer artifacts. This stability is likely attributed to our score-based loss, which provides a more stable training process than GAN-based methods.

Quantitative Results. Table 3 shows that our method achieves the lowest LPIPS and ArtFID scores, indicating superior content preservation and overall image quality. It also attains the second-best FID score, reflecting strong style adherence, second only to Scenimefy, which benefits from direct access to real anime data.

5.3. Ablation Studies

SMS Components. We systematically ablate each component in SMS to assess their contributions, with results shown in Figure 7. Without any identity regularization (*i.e.,* only the style matching objective), it successfully captures the overall Ghibli style but introduces noise and spurious details, such as background clutter in Figure 7(Row 1) (b)). Applying spectrum regularization L_{freq} improves structure integrity and color alignment with the source image but leaves some high-frequency artifacts (see Figure 7(c)). Semantic-aware gradient refinement R selectively stylizes important regions, reducing disharmonious details and better balancing style and content (see Figure 7(d)). However, due to the absence of direct regularization with the source image, content distortion and color mismatches still exist. Our full formulation, combining both components, achieves superior stylization with detailed textures, consistent colors, and strong content preservation (see Figure 7(e)). All the terms work together to improve the overall performance of the proposed style matching loss.

Table 4. **Quantitative ablation studies.**

Metric	w/o L_f, R	w/o R	w/o L_f	Full	Random t	Naïve t
LPIPS ↓	0.703	0.505	0.536	**0.326**	<u>0.389</u>	0.408
CFSD ↓	0.390	0.166	0.149	**0.102**	<u>0.107</u>	0.118
FID ↓	<u>14.500</u>	15.030	16.914	**13.089**	22.766	15.705
ArtFID ↓	26.403	24.132	27.514	**18.686**	32.936	<u>23.524</u>

Timestep Sampling Strategy. We compare our adaptive narrowing sampling strategy with random and naive linear decreasing sampling under identical settings. Random sampling causes blurriness (see Figure 7(f)), while naive linear decreasing sampling fails to preserve local identity, such as a red scard-like artifact appearing around the dog's neck (see Figure 7(g)). In contrast, our method produces sharp, detailed stylization while preserving content fidelity. The quantitative results in Table 4 clearly demonstrate the effectiveness of each proposed module. Additional ablation studies are provided in the supplementary material.

6. Conclusion

We propose Style Matching Score (SMS), an optimization method that formulates image stylization as a style distribution matching problem using style-specific LoRAs integrated with diffusion models. Our approach balances style transfer and content preservation through progressive spectrum regularization and semantic-aware gradient refinement. SMS also demonstrates its versatility by distilling style into lightweight feed-forward networks. Empirical results show its effectiveness in preserving content while adhering to target styles. Exploring SMS in broader parametric spaces, such as Neural Radiance Fields and 3D Gaussian Splatting can be interesting future work.

Acknowledgment. This study is supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2023). Yuxin Jiang is supported by the A*STAR ACIS Scholarship.

References

[1] Civit AI, Inc. https://civitai.com/. 3

[2] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. CartoonGAN: Generative adversarial networks for photo cartoonization. In *CVPR*, 2018. 2, 3

[3] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *CVPR*, 2024. 2, 3, 6, 7

[4] Martin Nicolas Everaert, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, and Radhakrishna Achanta. Diffusion in style. In *ICCV*, 2023. 1

[5] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 1, 2

[6] Feihong He, Gang Li, Mengyuan Zhang, Leilei Yan, Lingyu Si, Fanzhang Li, and Li Shen. FreeStyle: Free lunch for text-guided style transfer using diffusion models. *arXiv preprint arXiv:2401.15636*, 2024. 2, 3, 6, 7

[7] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta Denoising Score. In *ICCV*, 2023. 3, 6, 7, 8

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 7

[9] Lukas Höllein, Justin Johnson, and Matthias Nießner. StyleMesh: Style transfer for indoor 3d scene reconstructions. In *CVPR*, 2022. 1

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 2

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 3

[12] Yuxin Jiang, Liming Jiang, Shuai Yang, and Chen Change Loy. Scenimefy: learning to craft anime scene via semi-supervised image-to-image translation. In *ICCV*, 2023. 3, 7

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2

[14] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. PnP Inversion: Boosting diffusion-based editing with 3 lines of code. *ICLR*, 2024. 7

[15] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. *NeurIPS*, 2023. 7

[16] Juil Koo, Chanho Park, and Minhyuk Sung. Posterior Distillation Sampling. In *CVPR*, 2024. 3, 7, 8

[17] David McAllister, Songwei Ge, Jia-Bin Huang, David W. Jacobs, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Rethinking score distillation as a bridge between image distributions. *NeurIPS*, 2024. 3

[18] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Marcus A. Brubaker, Jonathan Kelly, Alex Levinshtein, Konstantinos G. Derpanis, and Igor Gilitschenski. Watch your steps: Local image and scene editing by text instructions. In *ECCV*, 2024. 2, 4

[19] Hyelin Nam, Gihyun Kwon, Geon Yeong Park, and Jong Chul Ye. Contrastive denoising score for text-guided latent diffusion image editing. In *CVPR*, 2024. 3

[20] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 3

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[23] Nataniel Ruiz, Yuanzhen Li, Neal Wadhwa, Yael Pritch, Michael Rubinstein, David E Jacobs, and Shlomi Fruchter. Magic insert: Style-aware drag-and-drop. *arXiv preprint arXiv:2407.02489*, 2024. 1

[24] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, 2022. 1

[25] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *ICCV*, 2021. 8

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[27] Yiren Song, Cheng Liu, and Mike Zheng Shou. Omniconsistency: Learning style-agnostic consistency from paired stylization data. *arXiv preprint arXiv:2505.18445*, 2025. 1

[28] Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. InstantStyle-Plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*, 2024. 2, 3, 6, 7

[29] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023. 2, 3

[30] Matthias Wright and Björn Ommer. ArtFID: Quantitative evaluation of neural style transfer. In *GCPR*, 2022. 7

[31] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 7

[32] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer. In *ICCV*, 2023. 2

[33] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024. 2, 3, 4

[34] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3

[35] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[36] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *CVPR*, 2023. 2, 3

[37] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2, 3