

RESEARCH ARTICLE

End-to-End Design of Webtoon-Style Portrait Stylization System for Real-World Demo Booth

SOJEONG KIM¹, BONG-SOO SOHN², AND JAESUNG LEE^{1,3}

¹Department of Artificial Intelligence, Chung-Ang University, Dongjak-gu, Seoul 06974, Republic of Korea

²School of Computer Science and Engineering, Chung-Ang University, Dongjak-gu, Seoul 06974, Republic of Korea

³AI/ML Innovation Research Center, Chung-Ang University, Dongjak-gu, Seoul 06974, Republic of Korea

Corresponding authors: Bong-Soo Sohn (bongbong@cau.ac.kr) and Jaesung Lee (curseor@cau.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by Korean Government (MSIT) [Artificial Intelligence Graduate School Program (Chung-Ang University)] under Grant 2021-0-01341, and in part by the Chung-Ang University Graduate Research Scholarship in 2023.

ABSTRACT Webtoon-style portrait stylization has gained prominence across various fields, including social media, virtual characters, and the Metaverse, with significant applications in demo booths for real-time audience interaction. However, achieving high-quality, real-time stylizations presents challenges, such as preserving facial identity while maintaining the webtoon aesthetics and ensuring robustness to diverse inputs. This study explores diffusion-based generative models for webtoon-style portrait stylization, focusing on real-world scenarios such as demo booths. We address challenges such as artifact reduction and adaptability to varying inputs, showcasing the potential of diffusion models to enhance interactive and artistic applications, particularly in demo booth settings.

INDEX TERMS Style transfer, image generation, image-to-image translation.

I. INTRODUCTION

Portrait stylization has had a notable impact on various fields, including social media filters, virtual character creation, animation production, advertising, and the Metaverse [1]. Among these applications, transforming portraits into specific artistic styles, such as the webtoon style, has gained significant attention owing to the growing popularity of webtoons, a type of digital comic widely published on online platforms. With the rapid digitization of comics, various industries are increasingly leveraging webtoon intellectual property to enhance user engagement and deliver immersive experiences [2], [3], [4]. For example, technologies that convert photos or videos into webtoon-style characters have become critical in creative industries, particularly in exhibition demos designed to engage and captivate diverse audiences [2].

As a common real-world scenario, demo booths serve as interactive platforms for showcasing cutting-edge technologies that often target parents and children. Webtoon-style

portrait stylization is an ideal medium for such demonstrations, enabling audiences to view photos transformed into personalized webtoon characters. This approach provides memorable and engaging experiences while demonstrating the artistic possibilities of AI-driven applications. However, several challenges must be addressed to ensure that high-quality webtoon-style portrait stylization is suitable for real-time demonstration applications. These include robustness to diverse input data, such as varying facial features and style images, and the ability to consistently produce visually compelling outputs. Previous approaches often failed to meet these demands, leading to artifacts and inconsistencies that do not meet user expectations [5], [6].

Recently, diffusion-based generative models have become powerful tools for style transfer tasks, offering the ability to generate detailed and high-quality outputs [7], [8], [9]. Although their potential has been demonstrated in general style transfer, applying these models to real-world scenarios such as demo booths requires careful consideration of the constraints described above. Previous studies have rarely addressed these constraints, leaving the challenge of optimizing these models for interactive and dynamic environments.

The associate editor coordinating the review of this manuscript and approving it for publication was Abdullah Ilyasu¹.

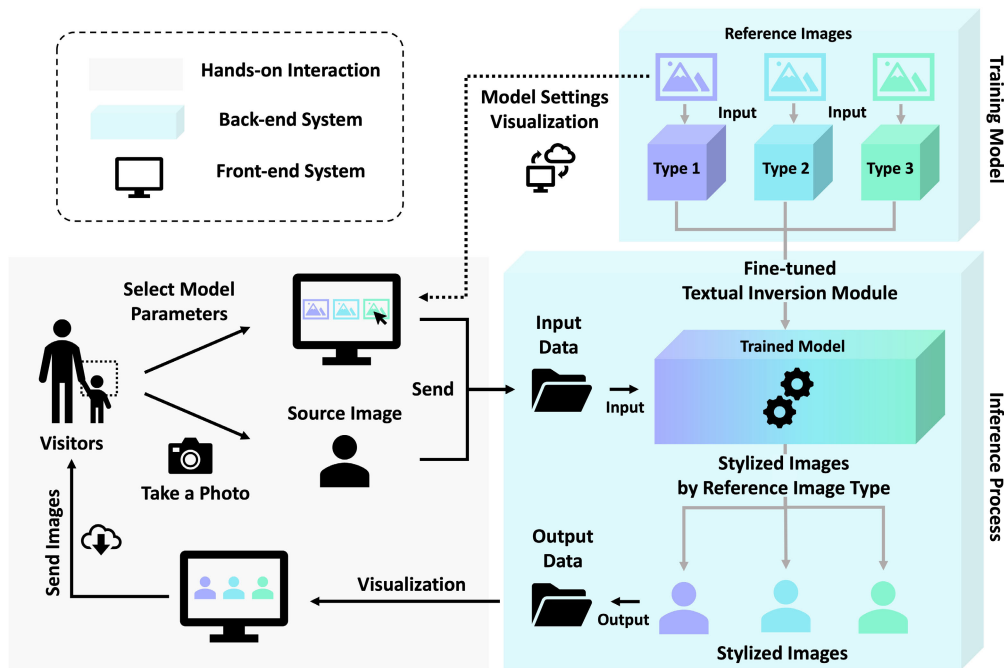


FIGURE 1. End-to-end system workflow: From user input to final result through front-end interaction and back-end processing.

In this study, we explore the use of diffusion-based generative models for webtoon-style portrait stylization designed for real-world scenarios, with a focus on demo booth platforms. Despite advancements in diffusion-based style transfer, challenges remain in achieving high-quality and harmonious stylization in interactive settings like demo booths, where it is essential to balance artistic style, ensure consistency across diverse inputs, and maintain practical inference speed for real-time applications.

To address these challenges, we propose a novel approach that blends visual and textual style features, enabling diffusion models to adaptively apply the reference style while maintaining the identity of the source image, resulting in a more harmonious stylization. Unlike previous methods relying solely on textual embeddings, our strategy prevents artifacts and distortions by integrating both modalities. In addition, we introduce Canny edge detection to guide the blending process, ensuring that the stylization preserves the semantics of the source image while integrating the desired artistic elements. This approach efficiently enhances style transfer performance, ensuring high stylization quality without significantly increasing inference time, which is particularly important for real-time applications.

In the experimental results, our proposed method demonstrates the best performance for real-time scenarios, such as demo booths. Balancing stylization performance and inference speed is essential in real-world style transfer applications. Our experimental results, conducted on 1,600 pairs, show that our proposed method is positioned closest

to the ideal point, demonstrating that it achieves the best balance between stylization performance and inference speed. Consequently, this well-balanced performance makes it the most suitable solution for real-time applications, such as demo booths, where efficiency and visual quality are essential.

In summary, the key contributions of this study are as

- **Optimized Real-Time Performance for Demo Booths:** Our method achieved the best trade-off between stylization performance and inference speed, making it the optimal solution for real-time scenarios.
- **Enhanced Style Transfer through Blended Embeddings:** We introduced a diffusion-based framework that fuses visual and textual style information, enabling more accurate and consistent style transfer across diverse image pairs.
- **Efficient Style Transfer with Edge-Guided Processing:** By incorporating Canny edge detection, we improved style transfer performance efficiently, achieving high-quality stylization without excessive computational overhead.

Building on these aspects, we achieve optimized performance in real-world webtoon-style portrait stylization. This method is particularly effective for real-time user interaction applications, such as demo booths. By ensuring high visual quality and efficiency, our proposed method contributes to creating an interactive and enjoyable experience, making visitors and their children happy at the demo booth.

TABLE 1. Challenges of webtoon-style portrait stylization.

Challenge	Description	Issues	Impact
Preservation of Facial Identity and Key Features	Maintaining facial identity and key features during stylization.	Distorted and misaligned facial features (eyes, lips, nose), resulting in unintended artifacts.	Undermines user satisfaction and reduces practical usability in demo booths.
Balancing Webtoon Aesthetic and Subject Identity	Achieving harmony between the source image and the webtoon aesthetic.	Over-stylization leads to distortion, while under-stylization fails to reflect the intended artistic style.	Directly impacts visual quality and coherence when catering to diverse audience preferences.
Robustness to Diverse Input Conditions	Ensuring consistency across diverse source images, including photos, cartoons, and diffusion-generated outputs.	Diverse input conditions can cause artifacts, oversaturation, or poorly defined style attributes.	Reduces real-world applicability in demo booths that accommodate diverse users and environments.

II. RELATED WORK

A. END-TO-END FRAMEWORKS

End-to-end frameworks have proven highly effective in various vision tasks, particularly in the fields of blind face restoration and image deblurring. In blind face restoration, several studies have integrated structural and generative priors into fully differentiable, end-to-end frameworks. GFP-GAN uses spatial feature transform layers for efficient one-pass restoration, SGPN combines 3DMM-based shape priors with StyleGAN2 manipulation, and FGPNet fuses pretrained GAN and landmark features within a U-Net [10], [11], [12]. More recently, diffusion-based methods such as IDM and PGDiff adopt end-to-end architectures that guide the restoration process through iterative refinement or partial structural priors, enabling robust performance without explicit degradation modeling [13], [14].

Similarly, in image deblurring, models such as DBLR-Net [15] and icDPM [16] follow an end-to-end design that directly maps blurred inputs to sharp outputs without intermediate processing. DBLRNet adopts an end-to-end architecture that jointly learns spatio-temporal features via 3D convolutions and integrates adversarial and content losses through a GAN framework [15]. At the same time, icDPM employs an end-to-end diffusion-based formulation conditioned on blurry inputs, enhanced by multiscale structure guidance without relying on explicit degradation modeling [16].

These studies demonstrate the effectiveness and flexibility of end-to-end frameworks in addressing complex vision tasks. By leveraging structural or diffusion-based priors, these approaches achieve high-quality results while ensuring adaptability and generalization across diverse degradation scenarios, thereby expanding their applicability to real-time settings.

B. IMAGE STYLE TRANSFER

Image style transfer, which can be effectively applied to portrait stylization, is achieved by analyzing statistical features or leveraging neural network feature correlations to modify a source image to reflect the style of a reference image [17]. This process is broadly classified into two approaches: optimization-based methods and feed-forward methods. Early studies defined style using the Gram matrix to capture deep feature correlations, balancing semantic similarity with

stylistic representation [17], [18]. This approach provided the foundational idea of using feature space statistics to measure style. However, it often lacked localized detail in stylized outputs. To address these limitations, a Markov random field (MRF)-based regularizer was proposed, replacing Gram matrix matching [19]. This method ensured local consistency in style patterns, maintaining coherent high-level feature representations throughout the image. The introduction of MRF regularization significantly improved the preservation of intricate style details.

Semantic-style transfer methods have been developed to enhance the alignment between style and content by targeting specific images regions [20], [21]. For example, a two-category semantic mask was used to guide the style transfer, enabling particular areas of the reference image to stylize the corresponding regions of the source image [20]. Another approach utilizes deep feature similarities to construct a correspondence map, facilitating the precise alignment of style elements with relevant regions in the source image [21]. These methods enable semantically meaningful visual transformations, thereby addressing the challenges of misaligned features in earlier techniques. Additionally, a significant advancement has been the utilization of pretrained VGG-19 for encoding image information [22]. This approach enables a hierarchical representation from low-level textures to high-level semantics, thereby allowing the accurate transfer of attributes, even between vastly different images.

In contrast, feedforward methods aim to directly optimize the style transfer network, thereby enabling faster and more efficient processing [23], [24]. One notable technique, AdaIN, adjusts the mean and variance of the source image features to match those of the reference image, effectively allowing a single model to support multiple predefined styles [23]. This approach achieves rapid style adaptation while retaining essential content features. Subsequent advancements included SANet, an attention-based method that replaces source image features with style features that are most closely matched at the patch level [24]. This mechanism allows for better style adherence in localized regions. Despite these improvements, feed-forward methods often struggle to produce seamless outputs at regional boundaries and manage complex style patterns.

Recently, diffusion models have become a powerful frameworks for high-quality image generation tasks,

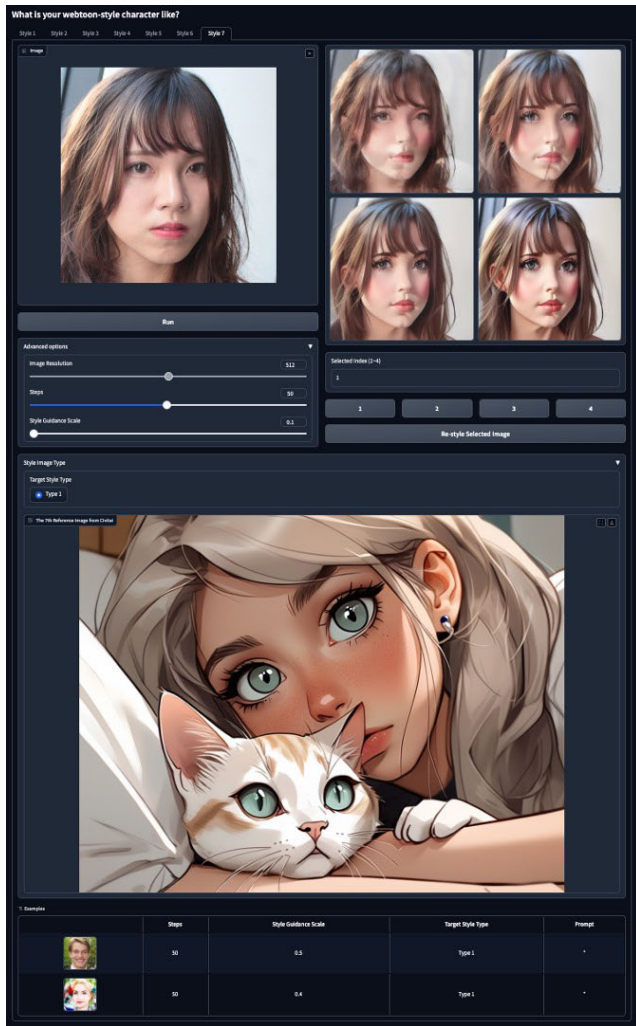


FIGURE 2. Overview of a UI screen with the input image on the top left and four corresponding webtoon-style output images on the top right, each generated with different style guidance scales—allowing users to select their preferred result—while reference style images from Civitai and example images are displayed at the bottom.

including portrait stylization [25], [26], [27], [28], [29]. These models iteratively refine outputs, resulting in detailed and consistent visual transformations. For instance, InST is an example-guided framework that uses diffusion models and attention-based textual inversion to transfer the style of a single reference image to the source images [26]. It preserves artistic attributes and content consistency while supporting text-based content control. Similarly, IP-Adapter introduces a decoupled cross-attention mechanism with dedicated layers for image features while preserving the original model [25]. This enables precise image-prompt-based generation, text-prompt compatibility, and seamless integration with adapters, such as ControlNet [30], for multimodal image generation. Similarly, Z-STAR adopts a training-free approach by leveraging generative priors from diffusion models [27]. It uses a rearranged attention mechanism in the latent space to fuse content and style from two input images, thereby

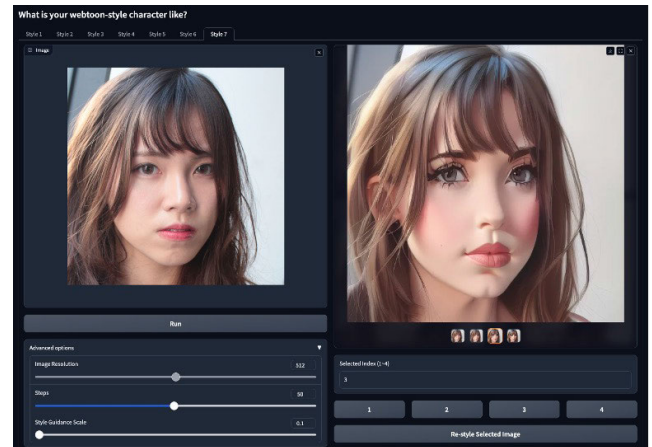


FIGURE 3. Top UI section displaying the input image and one upscaled stylized output image, selected from four results, each generated with a different style guidance scale.

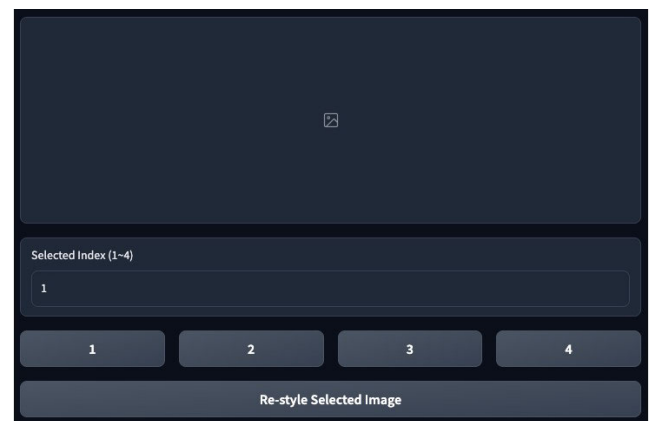


FIGURE 4. Right side of the top UI section allows users to select the index number of one of the four generated images for re-style transfer generation.

achieving a balance between preservation and adaptation without requiring retraining.

In contrast, DiffuseIT leverages an unconditional diffusion model for both text- and image-guided style transfers and maintains the structural content of the input image [28]. It employs DINO-ViT [31] for structural and style regularization, aligning the content and style using intermediate attention keys and [CLS] tokens. In addition, StyleID utilizes self-attention layers in large-scale pretrained diffusion models to transfer textures from a reference image while preserving the spatial structure of the source image [29]. It employs query preservation, attention temperature scaling, and initial latent AdaIN [23] to harmonize colors and refine textures, thereby achieving a state-of-the-art style transfer performance.

III. PROPOSED SYSTEM

In this section, we outline the key challenges of webtoon-style portrait stylization, such as preserving facial identity,

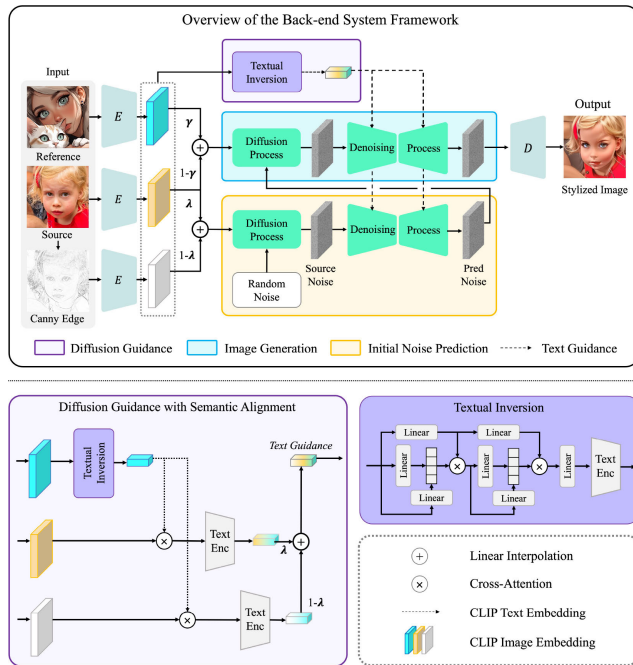


FIGURE 5. Overview of the back-end system framework and a detailed illustration of the diffusion guidance module.

balancing artistic aesthetics, and ensuring robustness to diverse input conditions. Next, we present the front-end design and back-end system algorithms developed to address these challenges and enable their real-time use in a demo booth. The overall end-to-end system workflow, illustrating the process from a visitor capturing a source image on-site, configuring the model parameters in the front-end system, submitting the input, and processing it through the pre-trained model to generate a stylized image, which is then visualized in the front-end system and ultimately delivered to the visitor, is depicted in Figure 1.

A. CHALLENGES OF WEBTOON-STYLE PORTRAIT STYLIZATION

Webtoon-style portrait stylization faces several challenges, such as preserving facial identity, balancing webtoon aesthetics with the subject's identity, and ensuring robustness to diverse input conditions, which can result in artifacts or incoherent outputs. These challenges are summarized in Table 1, which highlights their issues and impacts.

1) PRESERVATION OF FACIAL IDENTITY AND KEY FEATURES

One of the main challenges in webtoon-style portrait stylization is maintaining the facial identity and key features of the source image. During stylization, as the facial source image is transformed using the style features of the reference image, the facial features are often distorted or misaligned. Key features such as the eyes, lips, and nose may shift, and artifacts can appear when style features are applied to unintended regions of the face in the source image, resulting in degraded output quality. This is particularly problematic

for applications such as demo booths, which provide live demonstrations and personalized experiences. In real-world scenarios, user satisfaction is influenced not only by accuracy and speed but also by the preservation of facial identity. Particularly in interactive environments with parents and children, maintaining recognizable facial features is crucial for engagement. Excessive alteration of facial identity can undermine user satisfaction and reduce the practical usability of the systems.

2) BALANCING WEBTOON AESTHETIC AND SUBJECT IDENTITY

The distinct aesthetic of webtoon styles presents a challenge in generating a stylized output that achieves a harmonious balance between applying the artistic style features of the reference image and preserving the identity of the source subject. Overly strong stylization can lead to unnatural and severely distorted results from the original source image, while weaker stylization may fail to reflect the desired artistic style. The degree of this balance is critical, as it directly affects the visual quality and coherence of the output. An improper balance may result in outputs that are either overly artificial and detached from the original source image or fail to meet expectations for a stylized transformation. This challenge is particularly important when catering to diverse audiences with varying aesthetic preferences.

3) ROBUSTNESS TO DIVERSE INPUT CONDITIONS

Webtoon-style portrait stylization must be robust to diverse input conditions, including variations in lighting, skin tone, and different types of source images, such as photos, cartoons, and even diffusion model-generated outputs, which include intermediate stylized images in cascaded style transfer. Models that lack robustness may produce inconsistent results, leading to artifacts, oversaturation of the reference image, or poorly defined style attributes, particularly when handling diverse visual inputs and varying input conditions. This limits the real-world applicability, particularly in demo booths, which require seamless adaptation to diverse users and environments.

B. FRONT-END DESIGN AND INTERACTION

This section presents the front-end design developed to showcase a webtoon-style portrait stylization demo. Figures 2, 3, and 4 show sample user interface screens that illustrate the design and functionality. The interface was implemented using Gradio, which is a simplified and streamlined development tool. The key components and their functions are detailed in the following sections.

As shown in Figure 2, the overall layout of the Gradio UI is divided into three main sections. The upper section displays the source input image on the left and four corresponding stylized outputs on the right, each generated with a different style guidance scale, allowing users to select the result they prefer. The middle section features the reference style image

TABLE 2. Performance comparison of the embedding blending weight parameters: λ for diffusion guidance, ω for initial noise prediction, and γ for image generation, with individual LPIPS scores.

Embedding Blending Weights	λ	ω	γ	$\lambda(0.5) + \omega(0.5) + \gamma$
0.0	37.0628 $\pm$ 6.23	31.7706 $\pm$ 5.25	32.5962 $\pm$ 6.62	29.2809 $\pm$ 8.91
0.1	36.9531 $\pm$ 6.16	31.8766 $\pm$ 5.58	33.4947 $\pm$ 6.27	31.5067 $\pm$ 10.79
0.2	36.9082 $\pm$ 6.19	31.2372 $\pm$ 6.48	35.8851 $\pm$ 6.42	32.8462 $\pm$ 9.40
0.3	36.8367 $\pm$ 6.13	30.4920 $\pm$ 5.76	39.0671 $\pm$ 7.28	34.8204 $\pm$ 11.32
0.4	36.8340 $\pm$ 5.94	30.1448 $\pm$ 5.57	43.6810 $\pm$ 10.27	32.6942 $\pm$ 6.98
0.5	36.5731 $\pm$ 5.79	29.4696 $\pm$ 4.70	45.3801 $\pm$ 9.66	30.3086 $\pm$ 4.90
0.6	36.6604 $\pm$ 5.65	30.2348 $\pm$ 4.54	43.0208 $\pm$ 8.92	28.9634 $\pm$ 4.54
0.7	36.7247 $\pm$ 5.63	30.8587 $\pm$ 4.59	41.4216 $\pm$ 9.72	28.3306 $\pm$ 4.99
0.8	36.8688 $\pm$ 5.47	32.6668 $\pm$ 4.92	36.7496 $\pm$ 6.84	28.4634 $\pm$ 4.96
0.9	36.8250 $\pm$ 5.28	34.2919 $\pm$ 5.39	35.7008 $\pm$ 5.92	28.4207 $\pm$ 5.12
1.0	36.8296 $\pm$ 5.20	35.5233 $\pm$ 5.05	35.5233 $\pm$ 5.05	29.0056 $\pm$ 5.21

used for style transfer, while the lower section presents a range of example source images for users to explore.

Focusing on the upper section, as shown in Figure 3, the left panel accepts the source input image and includes adjustable parameters such as image resolution, DDIM steps, and style guidance. Increasing the number of DDIM steps improves the quality and detail of the output image, but also increases the processing time required to generate the output. Similarly, the style guidance parameter controls how strongly the reference style is applied to the output image. Higher values resulted in a stronger application of the reference style, whereas lower values retained more characteristics of the original source image. Upon clicking the Run button, the system generates four output images on the right panel by adjusting the style guidance scale in increments up to +0.3. This allows users to compare and select the result they prefer most. In addition, the interface offers a re-style transfer generation function if none of the four images is satisfactory. As shown in Figure 4, users can select the index number of any of the four images, confirm their selection by the number above the button, and click the “Re-style Selected Image” button to regenerate the output. This design focuses on user-centered interaction, helping users find the most satisfying result.

C. BACK-END SYSTEM

The backend system is designed with a focus on a balanced and harmonious style transfer, ensuring that the semantics of the source image are not severely distorted when transferring style attributes from the reference image. We adopt InST as the base model and use linear interpolation to integrate the Canny edge map into the following two modules. The overall framework of the back-end system is illustrated in Figure 5.

1) DIFFUSION GUIDANCE WITH SEMANTIC ALIGNMENT

In the back-end system, diffusion guidance combines semantic embeddings from the Canny edge map through cross-attention layers. The combined embeddings were

processed as

$$v_{\text{new}} = (1 - \lambda) \cdot \text{Attention}(Q, K^c, V^c) + \lambda \cdot \text{Attention}(Q, K^s, V^s), \quad (1)$$

where K^c, V^c are from the Canny edge map embeddings, and K^s, V^s are from the source image embeddings. The parameter λ controls the blending intensity between the two embeddings. Table 2 shows the effect of different λ values on performance.

2) INITIAL NOISE PREDICTION

The initial noise prediction module incorporates semantic information from the source image using the Canny edge map. The embedding fusion of the source image and its Canny edge map is performed as

$$z_f = (1 - \omega) \cdot E(I_c) + \omega \cdot E(I_s), \quad (2)$$

where $E(I_c)$ and $E(I_s)$ are the CLIP embeddings of the Canny edge map and the source image, respectively. The parameter ω controls the blending intensity between the two embeddings. Table 2 compares the performance under different ω values.

Using fused embedding z_f , the initial noise map at time step t is computed as

$$z_t = z_f + \epsilon_t, \quad (3)$$

where ϵ_t denotes the added Gaussian noise. The predicted noise at step t is calculated as

$$\epsilon_t^{\text{pred}} = (z_{t-1} - \mu_T(z_t, t, v_{\text{new}})) \sigma_t, \quad (4)$$

3) IMAGE GENERATION

The fused embedding z_t is linearly interpolated between the embeddings of the reference image I_r and the source image I_s , with the predicted noise ϵ_t^{pred} added as

$$z_t = (1 - \gamma) \cdot E(I_r) + \gamma \cdot E(I_s) + \epsilon_t^{\text{pred}}, \quad (5)$$

where γ adjusts the balance between the source and reference images. Table 2 presents the performance comparison for different γ values.

Classifier-free guidance refines the process, enabling control over stylization intensity through the guidance scale (s), defined as

$$\hat{\epsilon}_\theta(z_t, t, v_{\text{new}}) = s \cdot \epsilon_\theta(z_t, t, v_{\text{new}}) + (1 - s) \cdot \epsilon_\theta(z_t, t, v_\emptyset), \quad (6)$$

where $v_\emptyset$ represents the null text embedding.

4) EFFICIENT SAMPLING WITH DDIM

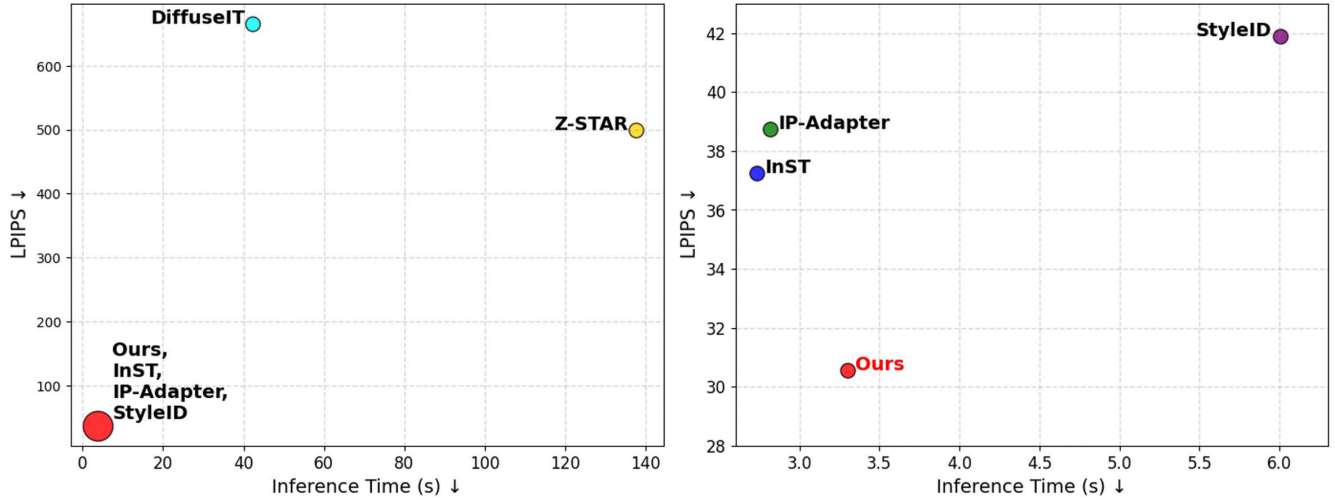
The back-end system employs the DDIM sampler to reduce the number of sampling iterations for high-quality image generation. The final denoised latent code $\hat{z}_0$ is computed as

$$\hat{z}_0 = \frac{z_t - \sqrt{1 - \bar{\alpha}_t} \cdot \hat{\epsilon}_\theta(z_t, t, v_{\text{new}})}{\sqrt{\bar{\alpha}_t}}, \quad (7)$$

where $\bar{\alpha}_t$ is the noise scaling factor.

TABLE 3. Quantitative comparison of the proposed method and other models under the ArtFID framework using three perceptual distance metrics-LPIPS, MS-SSIM, and CSIM-along with user score and inference time (↑/↓ symbol means larger/smaller values are better, respectively).

Model	LPIPS↓	MS-SSIM↓	CSIM↓	User Score↑ (LPIPS)	Inference Time (Sec.)↓ (LPIPS)
Ours	30.5643 ± 5.04	24.2075 ± 4.32	106.3852 ± 21.63	42.1%	3.3004
InST	37.2566 ± 5.04	27.0406 ± 3.86	121.9447 ± 23.74	11.2%	2.7345
DiffuseIT	666.0543 ± 43.34	484.4775 ± 32.52	7361.4451 ± 3821.98	6.9%	42.3431
IP-Adapter	38.7595 ± 7.71	27.3236 ± 5.33	108.4821 ± 29.56	13.4%	2.8161
StyleID	41.8970 ± 6.03	30.5459 ± 4.24	206.7092 ± 93.55	26.2%	6.0034
Z-STAR	499.7536 ± 77.36	356.8058 ± 55.62	2209.2775 ± 884.50	0.3%	137.6689

**FIGURE 6.** Comparison of style transfer model performance: LPIPS vs. inference time (Left: full range, right: zoomed-in view).

IV. EXPERIMENTAL RESULTS

In this section, we conduct experiments to analyze how diffusion-based models address the challenges outlined in Section III-A, focusing on preserving facial identity and integrating webtoon aesthetics.

A. EXPERIMENTAL SETTINGS

In our experiments, we used Stable Diffusion v1-4 with default settings (learning rate 1e-4, resolution 512 × 512) on an NVIDIA RTX 3090. We used the FFHQ [32] as the source images and shared images on Civitai as the reference images using Stable Diffusion-based models.

To evaluate the performance of style transfer, we adopt the ArtFID metric to jointly evaluate both source preservation and reference style matching in stylized images. The formulation of ArtFID is

$$\text{ArtFID}(X_g, X_{\text{src}}, X_{\text{ref}}) = \left(1 + \frac{1}{N} \sum_{i=1}^N d(X_{\text{src}}^{(i)}, X_g^{(i)})\right) + \text{FID}(X_{\text{ref}}, X_g). \quad (8)$$

Here, X_{src} , X_{ref} and X_g denote the sets of source, reference style, and generated stylized images, respectively. The first term measures content distortion via a perceptual distance metric $d(\cdot)$, and the second quantifies style consistency using

**FIGURE 7.** Stylized images generated from the FFHQ dataset with webtoon-style reference images using our proposed method. The image in the first row and column is used as a source and reference, respectively.

FID. We used three metrics for $d(\cdot)$: Learned Perceptual Image Patch Similarity (LPIPS) [33], Multiscale structural Similarity (MS-SSIM) loss [34], and Cosine similarity loss (CSIM), which capture perceptual similarity, structural consistency, and semantic identity preservation, respectively. LPIPS can be defined as

$$\text{LPIPS}(x, y) = \sum_l \frac{1}{H_l W_l} \sum_{h, w} \|w_l \odot (\hat{f}_l^x(h, w) - \hat{f}_l^y(h, w))\|_2^2, \quad (9)$$

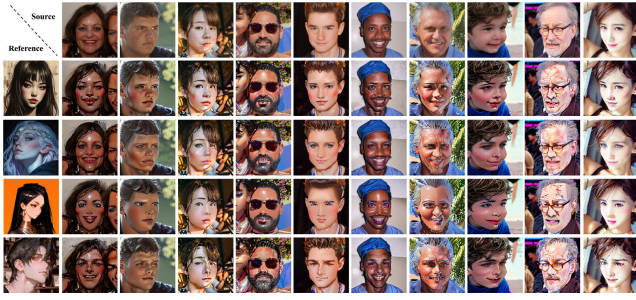


FIGURE 8. Stylized images generated from the FFHQ dataset with webtoon-style reference images using the InST method. The image in the first row and column is used as a source and reference, respectively.

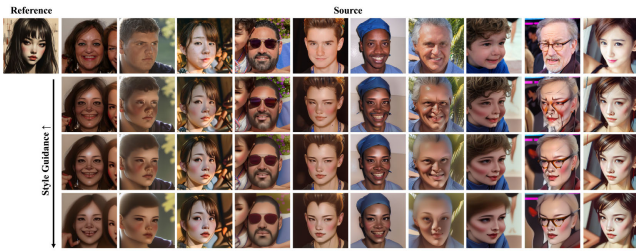


FIGURE 9. Stylized images generated from the FFHQ dataset using webtoon-style reference images, with increasing style guidance applied. Images further down represent results with higher style guidance values, generated using our proposed method.

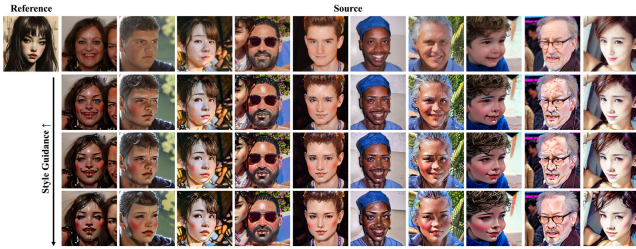


FIGURE 10. Stylized images generated from the FFHQ dataset using webtoon-style reference images, with increasing style guidance applied. Images further down represent results with higher style guidance values, generated using the InST method.

where $\hat{f}_l^x$ and $\hat{f}_l^y$ denote the normalized deep features from layer l of a pretrained network for images x and y , respectively, and w_l is a learned channel-wise weight. Next, MS-SSIM loss can be calculated as

$$\begin{aligned} \text{MS-SSIM}(x, y) &= [I_M(x, y)]^{\alpha_M} \cdot \prod_{j=1}^M [c_j(x, y)]^{\beta_j} [s_j(x, y)]^{\gamma_j}, \quad (10) \end{aligned}$$

where I_M is the luminance comparison at scale M , and c_j and s_j are the contrast and structure comparisons at scale j . Lastly, Cosine similarity loss can be calculated as

$$\text{CSIM}(x, y) = 1 - \frac{x \cdot y}{|x| |y|}, \quad (11)$$

where x and y are ArcFace [35] embeddings represent high-level identity features for evaluating identity

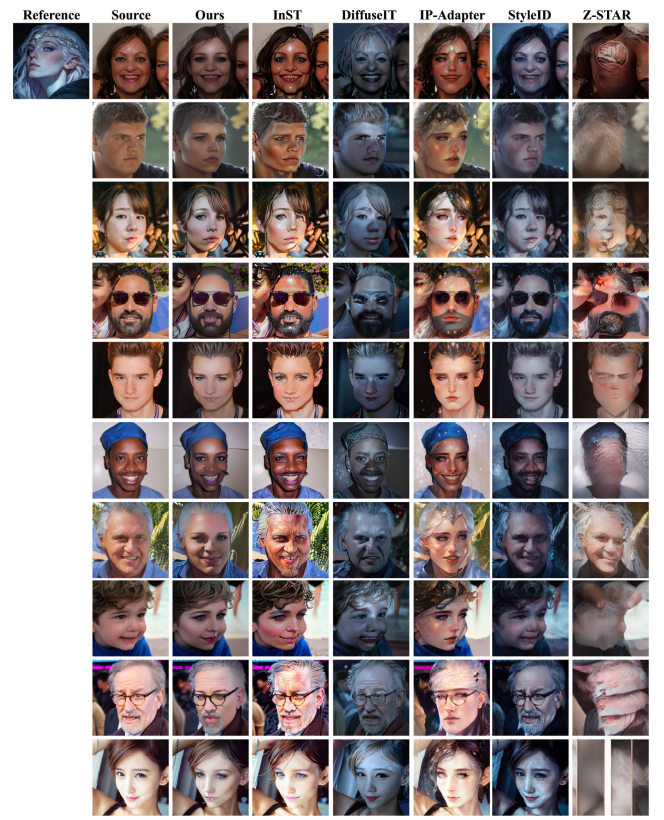


FIGURE 11. Stylized images generated from the FFHQ dataset using a reference style image of type 1, created by generative AI from Civitai, applying different diffusion-based style transfer models.

TABLE 4. Comparison results of generated images with different style guidance scales, increasing from 0.1 to 0.9 in increments of 0.1, using our proposed method.

Style Guidance Scale	LPIS $\downarrow$	User Score $\uparrow$
0.1	33.3321 $\pm$ 3.06	1.2%
0.2	34.8679 $\pm$ 2.98	10.4%
0.3	32.8368 $\pm$ 3.88	22.4%
0.4	28.9018 $\pm$ 4.30	24.2%
0.5	26.5107 $\pm$ 3.38	21.1%
0.6	27.6893 $\pm$ 4.36	19.5%
0.7	29.4123 $\pm$ 5.68	1.1%
0.8	30.3429 $\pm$ 6.57	0.1%
0.9	31.8602 $\pm$ 8.74	0.0%

preservation. ArcFace is a face recognition model with highly discriminative identity features, making it a reasonable choice for evaluating face identity consistency in generated images. Subsequently, throughout the experiments, those three ArtFID-based metrics will be referred to as LPIPS, MS-SSIM, and CSIM for brevity.

1) PERFORMANCE COMPARISON ACROSS EMBEDDING BLENDING WEIGHTS

We measured LPIPS scores for each embedding blending weight using 100 randomly sampled reference-source pairs,

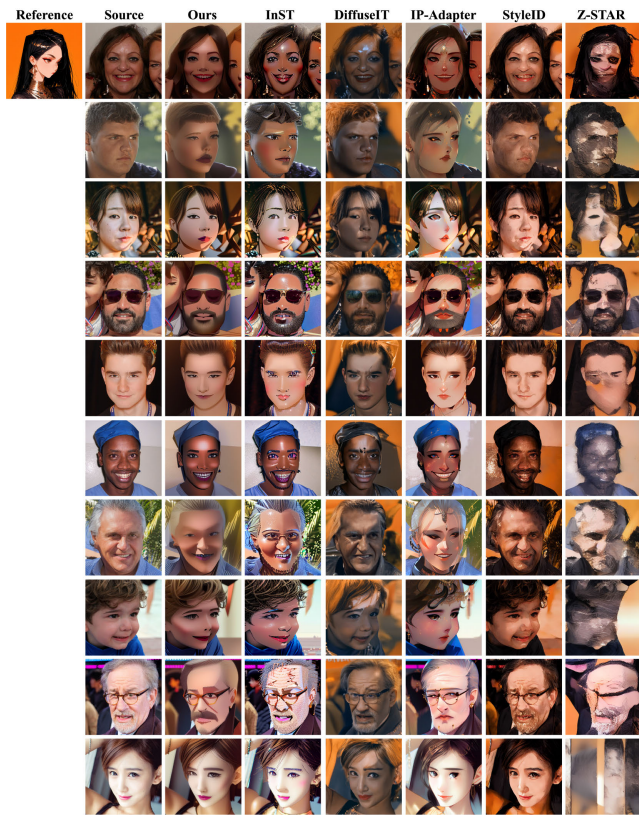


FIGURE 12. Stylized images generated from the FFHQ dataset using a reference style image of type 2, created by generative AI from Civitai, applying different diffusion-based style transfer models.

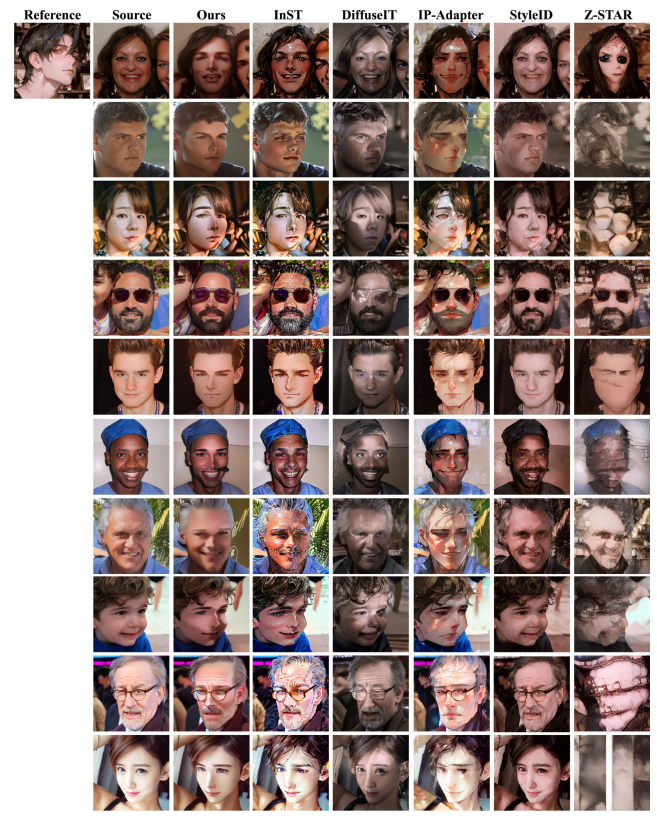


FIGURE 13. Stylized images generated from the FFHQ dataset using a reference style image of type 3, created by generative AI from Civitai, applying different diffusion-based style transfer models.

generated from 5 webtoon-style reference images and 20 source images from the FFHQ dataset. The quantitative results are presented in Table 2. In the columns labeled λ , ω , and γ , each parameter was individually varied according to the embedding blending weight values, while the other parameters were fixed to their default values as used in the InST model. For the $\lambda(0.5) + \omega(0.5) + \gamma$ column, the λ and ω values were fixed at 0.5, corresponding to the lowest LPIPS values observed in earlier settings, while only γ was varied to evaluate its impact across different embedding blending weights.

As shown in Table 2, the lowest LPIPS scores for both λ and ω are observed when their embedding blending weights are set to 0.5. Although the lowest LPIPS for γ occurs at 0.0 when it is varied alone, the best performance is achieved when γ is set to 0.7 while keeping λ and ω fixed at 0.5. Based on these results, it seems reasonable to set the default embedding blending weights for our proposed method to $\lambda = 0.5$, $\omega = 0.5$, and $\gamma = 0.7$.

2) PERFORMANCE COMPARISON ACROSS MODELS

Table 3 presents the evaluation metrics measured on a total of 1,600 stylized images, generated using 8 different styles, with 200 stylized images per style for each model.

We also measured inference time to assess efficiency in real-time applications, and a user-study-based user score to reflect quantitative performance. The user study involved 23 participants (16 males, 7 females) with an average age of 32.9 ± 12.4 years, who assessed 50 randomly selected source-reference pairs, resulting in 1,150 votes. The user score was expressed as a percentage and calculated based on the proportion of votes received relative to the total number of votes.

As shown in Table 3, the experimental results indicate that our method achieves the best performance in LPIPS, MS-SSIM, and CSIM within the ArtFID framework, demonstrating perceptually natural and identity-preserving style transfer while maintaining artistic coherence. The user score further supports this, as our model received the highest rating compared to the other models, validating its qualitative superiority in terms of visual appeal and user preference. In addition, in terms of inference time, Ours, InST, and IP-Adapter complete inference in around 3 seconds, making them the most suitable models for real-time use in demo booths.

To provide a more intuitive comparison, we plotted the inference time on the x-axis and LPIPS on the y-axis, as shown in Figure 6. The left graph illustrates the performance of all the style transfer models, whereas

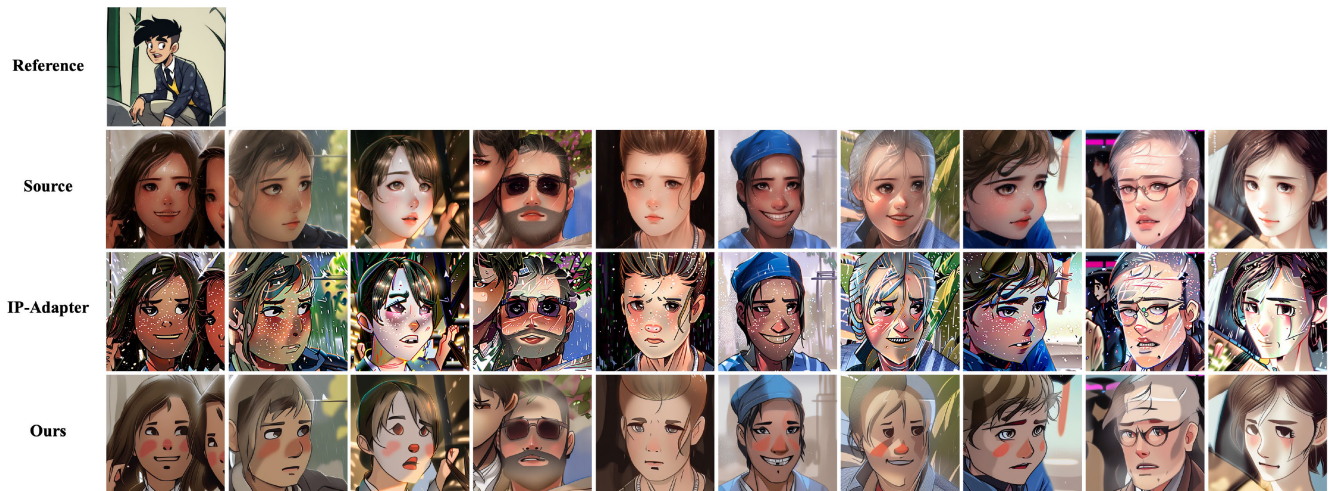


FIGURE 14. Style transfer results using IP-Adapter stylized images as source inputs for reference image 4.

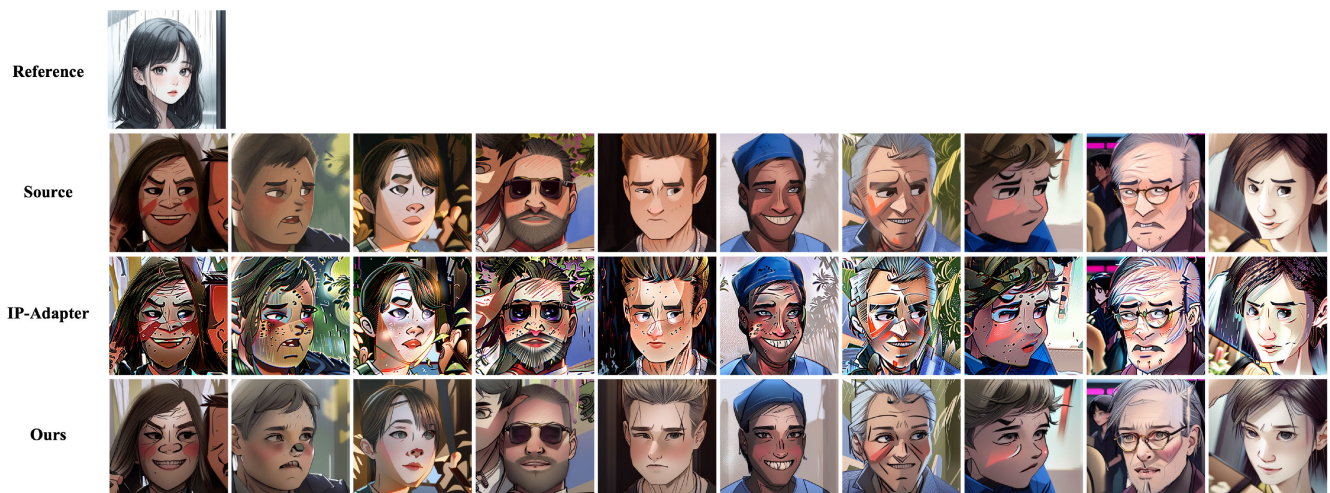


FIGURE 15. Style transfer results using IP-Adapter stylized images as source inputs for reference image 5.

the right graph zooms in on a specific range for a more detailed comparison. In real-world scenarios, both accuracy and inference time are crucial factors. While other systems have their own advantages, Figure 6 shows that our system is closest to the ideal point, achieving the best trade-off between style transfer performance and inference speed. Consequently, our model was demonstrated to be the most suitable model for real-time demo applications and was ultimately selected for the actual demonstration showcase.

3) PRESERVATION OF FACIAL IDENTITY AND KEY FEATURES

In this section, we conducted qualitative experiments that compared various diffusion-based models, evaluating their ability to preserve facial identity and maintain the structural integrity of key features across a range of face sample inputs. The stylized images in Figures 7 to 13 were generated from different reference images using various diffusion-based

models [25], [26], [27], [28], [29]. As shown in the figures, in terms of preserving the facial identity, the model named Z-STAR [27] exhibited the most significant distortions, whereas the other models maintained the overall structure of the subject with different levels of detail preservation. This can be attributed to its tendency to overfit the style of the reference image, leading to the loss of key facial features and identity consistency. Notably, the StyleID [29] model demonstrated the best preservation of detailed features from the original source image. However, the overall color scheme and skin tone of the reference image are strongly transferred to the stylized images, resulting in noticeable changes in the skin tone of the source images.

In contrast, the DiffuseIT [28] model largely disregards the color distribution of the source image, directly transferring the color of the reference image scheme regardless of the object placement in the source image. Consequently, the overall color scheme of the stylized image closely resembles that of the reference image while maintaining its distribution

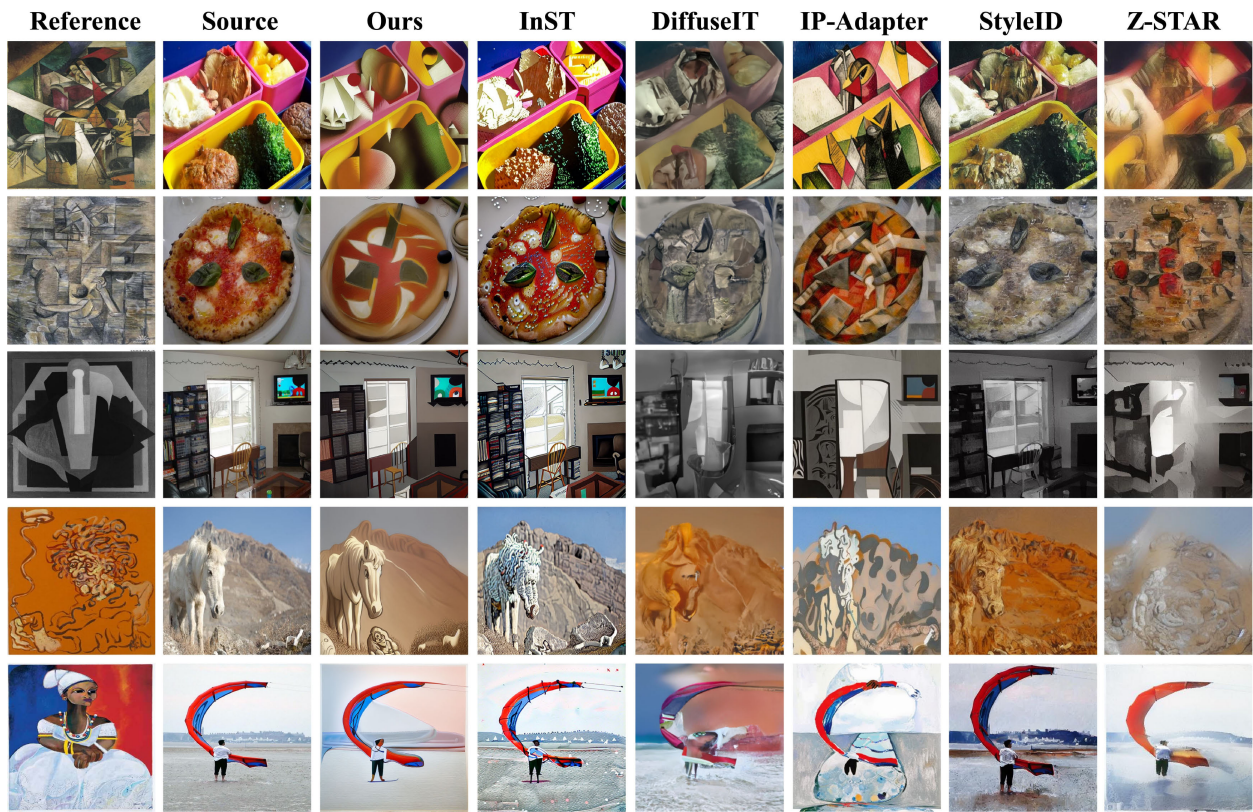


FIGURE 16. Qualitative comparisons with conventional style transfer models using five source–reference pairs, where reference images are sampled from the WikiArt dataset and source images from MS-COCO 2014.

and tone. For the IP-Adapter [25] model, the color and overall structure of the source image were well-preserved. However, the fine details of the facial features were significantly transformed to match those of the reference images. On the other hand, the InST [26] model showed better preservation of the original facial features from the source image compared to IP-Adapter. Nevertheless, it still introduced unnecessary artifacts, as shown in Figure 11, and exhibited more distortion in the detailed features than StyleID.

To further enhance the preservation of detailed structures in the source image, we incorporated semantic shape information using a Canny edge map during the inference stage, as described in Section III-C. The generated examples in the “Ours” column illustrate this approach. Figures 7 and 9 show the stylized images generated for each reference and source image pair, respectively, using this method. Overall, incorporating the Canny edge map improved the preservation of fine details compared with the original InST model, as shown in Figures 11, 12, and 13. However, in some cases, its performance in terms of detail preservation falls short of that achieved by the StyleID model.

4) BALANCING WEBTOON AESTHETIC AND SUBJECT IDENTITY

In this section, we examine how style guidance parameters influence this balance and evaluate their impact on achieving

TABLE 5. Comparison results between our method and conventional style transfer models using an additional 500 reference–source pairs from the WikiArt and MSCOCO 2014 datasets. (↑/↓ symbol means larger/smaller values are better, respectively).

Models	LPIS↓	User Score↑
Ours	82.9125 ± 27.81	37.3%
InST	87.6064 ± 36.20	11.8%
DiffuseIT	492.5427 ± 47.11	5.4%
IP-Adapter	83.1861 ± 28.82	19.6%
StyleID	86.6770 ± 26.58	25.7%
Z-STAR	365.5884 ± 42.36	0.2%

high-quality stylized outputs. In Figure 9, stylized images generated using the proposed method are shown with varying style guidance scales. From top to bottom, each row corresponds to one of the three highest user score values identified in Table 4, specifically at style guidance scales of 0.3, 0.4, and 0.5. Overall, as the style-guidance scale increased, the facial features of the original source image changed, transforming them to resemble elements of the reference image. This effect is particularly evident in the eighth column, where an increase in style guidance causes the teeth visible in the original source image to disappear, thereby transforming the output to match the lip shape of the reference image. However, by incorporating the Canny edge map during the inference stage, semantic shape preservation

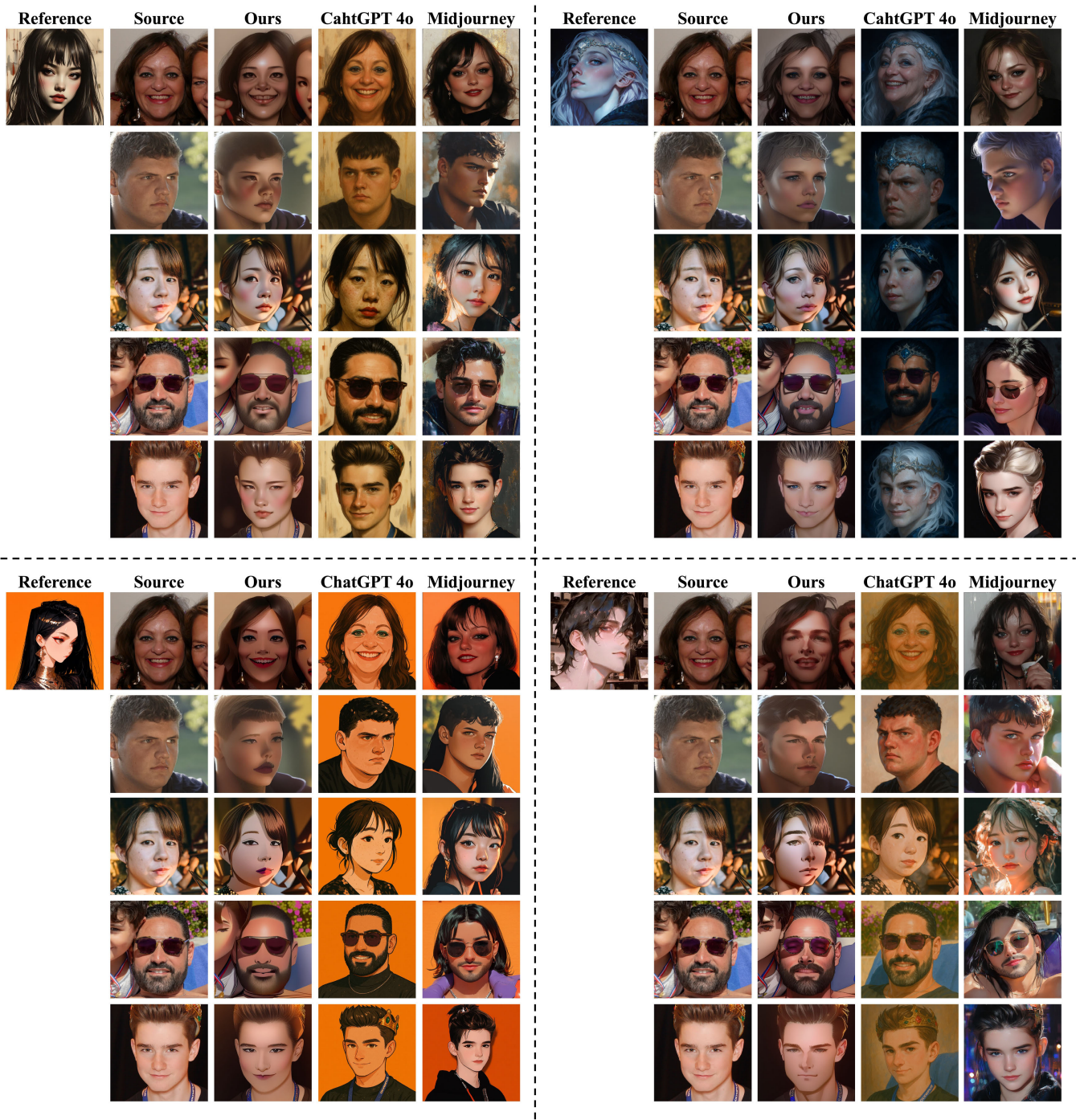


FIGURE 17. Comparison between our proposed method and existing commercial and real-time stylization systems (GPT-4o and Midjourney) using four different webtoon-style reference images.

is improved, resulting in less distortion of the original subject compared to the original InST method, even as the style guidance scale increases, as shown in Figures 9 and 10.

In addition, as shown in Table 4, the ablation study quantitatively evaluates the performance of stylized images generated by the proposed method across different style guidance scales. Style transfer was performed using 20 images from the FFHQ dataset for each of 8 different

reference style images, resulting in a total of 160 stylized outputs. The results show that the overall average error rate, measured by the LPIPS metric, remains below 8%, with a standard deviation of less than 3, indicating consistent performance across style guidance scales. Notably, user scores were relatively evenly distributed across the range of 0.3 to 0.6, suggesting that individual preferences vary. These findings highlight the importance of presenting multiple

stylized outputs with different style guidance scales during the demo, allowing users to compare and choose their preferred result. The overall UI interface reflecting this approach is illustrated in Figure 2.

5) ROBUSTNESS TO DIVERSE INPUT CONDITIONS

In this section, to ensure robustness to diverse input conditions, especially with diverse source images, we explored a cascaded style transfer approach, where stylized images are used as source images for further style transfer with new source-reference pairs. By analyzing the impact of using stylized images as source inputs, we aimed to assess the style consistency and overall visual quality across different source-reference pairs.

As seen in Figures 14 and 15, when performing style transfer using stylized images generated by IP-Adapter as source images, we observed the appearance of unwanted artifacts, such as white or black specks across the images. These artifacts appear to originate from the style effects embedded in the stylized images and were likely treated as part of the semantic content that needed to be preserved during the style transfer process, leading to their unintended presence in the final outputs. By contrast, when using our proposed method, the artifacts observed in the previous case were not present, resulting in more harmonious outputs. This suggests that the model can effectively handle not only photo-style source images but also various stylized images, demonstrating its versatility in style transfer applications.

Furthermore, we conducted an ablation study using additional reference-source dataset pairs, specifically employing WikiArt [36] as the reference dataset and MSCOCO 2014 [37] as the source dataset, resulting in a total of 500 generated images. As shown in Table 5, our proposed method achieved the best performance regarding both the LPIPS metric and user score. Stylized images generated by each model are presented in Figure 16. These results further support the robustness of our method across different datasets.

6) COMPARISON TO COMMERCIAL AND REAL-TIME STYLIZATION SYSTEMS

As shown in Figure 17, our proposed method generally maintains the color tone of the source image while applying other stylistic elements from the reference images. In contrast, ChatGPT 4o often transfers the color tone and specific details from the reference images, such as the hairstyle or hair accessories of the character, onto the source image. This is particularly noticeable in the case of the top-right reference image. As a result, while the overall facial features remain somewhat similar to the source, the composition and positioning of the character often change significantly. Also, in the case of the bottom-right reference image, the generated output reflects a style that differs notably from that of the reference. On the other hand, Midjourney shows significant alterations in both facial features and character composition, often deviating substantially from the source image. Except

for the first reference case, the stylization from the remaining reference images appears to be poorly transferred.

These results suggest that our proposed method shows strong potential for practical applicability and competitiveness compared to existing commercial stylization systems, as it effectively preserves key elements of the source image, such as color tone, facial structure, background, and composition, while applying relevant stylistic features from the reference image.

V. CONCLUSION

In this paper, we conducted a series of experiments to evaluate the strengths and weaknesses of different diffusion-based style transfer methods in addressing the aforementioned issues. Our findings highlight key considerations for enhancing webtoon-style portrait stylization in real-world scenarios where demo booths are prevalent interactive platforms. The insights gained from this research provide a foundation for developing more robust and versatile models capable of meeting the practical demands of real-world scenarios, such as demo booth settings.

Existing Stable Diffusion-based style transfer methods mainly use photo-style source images, limiting cross-domain applicability and generalization to arbitrary source-reference pairs. Therefore, future work will explore cascaded style transfer for cross-domain scenarios, enabling artistic style transfer to and from nonphotorealistic images (e.g., sketches and cartoons) for greater flexibility. In addition, we plan to extend the cascaded style transfer approach to cross-domain style transfer, applying artistic styles to nonphotorealistic images (e.g., sketches and cartoons), or vice versa, to further explore its adaptability and effectiveness across diverse visual domains. This extension can enhance the real-time applications in demo booths, offering users a more immersive and interactive experience by enabling dynamic style transformations across different media types.

REFERENCES

- [1] M. Zawish, F. A. Dharejo, S. A. Khowaja, S. Raza, S. Davy, K. Dev, and P. Bellavista, "AI and 6G into the metaverse: Fundamentals, challenges and future research trends," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 730–778, 2024.
- [2] J. Back, S. Kim, and N. Ahn, "WebtoonMe: A data-centric approach for full-body portrait stylization," in *Proc. SIGGRAPH Asia Tech. Commun.*, Dec. 2022, pp. 1–4.
- [3] K. Yu, H. Kim, J. Kim, C. Chun, and P. Kim, "A study on generating webtoons using multilingual text-to-image models," *Appl. Sci.*, vol. 13, no. 12, p. 7278, Jun. 2023.
- [4] S. Kim, S. Kim, and S.-H. Nam, "A framework for portrait stylization with skin-tone awareness and nudity identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2024, pp. 3660–3664.
- [5] M. J. Chong and D. Forsyth, "JoJoGAN: One shot face stylization," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, Jan. 2022, pp. 128–152.
- [6] S. Kim, C. Gwak, D. Kim, K. Lee, J. Back, N. Ahn, and D. Kim, "Cross-domain style mixing for face cartoonization," 2022, *arXiv:2205.12450*.
- [7] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22500–22510.

- [8] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," 2022, *arXiv:2208.01618*.
- [9] B. Cheng, Z. Liu, Y. Peng, and Y. Lin, "General image-to-image translation with one-shot image guidance," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 22679–22689.
- [10] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9164–9174.
- [11] F. Zhu, J. Zhu, W. Chu, X. Zhang, X. Ji, C. Wang, and Y. Tai, "Blind face restoration via integrating face shape and generative priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7652–7661.
- [12] H. Shao, J. Bai, R. Wu, J. Jiang, and H. Liang, "FGPNet: A weakly supervised fine-grained 3D point clouds classification network," *Pattern Recognit.*, vol. 139, Mar. 2023, Art. no. 109509.
- [13] Y. Zhao, T. Hou, Y.-C. Su, X. Jia, Y. Li, and M. Grundmann, "Towards authentic face restoration with iterative diffusion models and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7278–7288.
- [14] P. Yang, S. Zhou, Q. Tao, and C. C. Loy, "PGDiff: Guiding diffusion models for versatile face restoration via partial guidance," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2023, pp. 32194–32214.
- [15] K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu, and H. Li, "Adversarial spatio-temporal learning for video deblurring," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 291–301, Jan. 2019.
- [16] M. Ren, M. Delbracio, H. Talebi, G. Gerig, and P. Milanfar, "Multiscale structure guided diffusion for image deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 10687–10699.
- [17] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.
- [18] S. Schreiber, J. Geldenhuys, and H. de Villiers, "Texture synthesis using convolutional neural networks with long-range consistency and spectral constraints," in *Proc. Pattern Recognit. Assoc. South Afr. Robot. Mechatronics Int. Conf. (PRASA-RobMech)*, Nov. 2016, pp. 1–6.
- [19] C. Li and M. Wand, "Combining Markov random fields and convolutional neural networks for image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2479–2486.
- [20] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman, "Controlling perceptual factors in neural style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3730–3738.
- [21] K. Aberman, J. Liao, M. Shi, D. Lischinski, B. Chen, and D. Cohen-Or, "Neural best-buddies: Sparse cross-domain correspondence," *ACM Trans. Graph. (TOG)*, vol. 37, no. 4, p. 69, Jul. 2018.
- [22] J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang, "Visual attribute transfer through deep image analogy," *ACM Trans. Graph.*, vol. 36, no. 4, p. 120, 2017.
- [23] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.
- [24] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5873–5881.
- [25] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "IP-adapter: Text compatible image prompt adapter for text-to-image diffusion models," 2023, *arXiv:2308.06721*.
- [26] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu, "Inversion-based style transfer with diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 10146–10156.
- [27] Y. Deng, X. He, F. Tang, and W. Dong, "Z*: Zero-shot style transfer via attention reweighting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 34, Jun. 2024, pp. 6934–6944.
- [28] G. Kwon and J. C. Ye, "Diffusion-based image translation using disentangled style and content representation," 2022, *arXiv:2209.15264*.
- [29] J. Chung, S. Hyun, and J.-P. Heo, "Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 8795–8805.
- [30] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3813–3824.
- [31] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640.
- [32] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2018, *arXiv:1812.04948*.
- [33] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
- [34] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, pp. 1398–1402.
- [35] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022.
- [36] F. Phillips and B. Mackintosh, "Wiki art gallery, Inc.: A case for critical thinking," *Issues Accounting Educ.*, vol. 26, no. 3, pp. 593–608, Aug. 2011.
- [37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2014, pp. 740–755.



SOJEONG KIM is currently pursuing the M.S. degree with the Department of Artificial Intelligence, Chung-Ang University, Seoul, South Korea. Her research interests include computer vision and example-guided style transfer.



BONG-SOO SOHN received the B.S., M.S., and Ph.D. degrees in computer sciences from the University of Texas at Austin, in 1999, 2001, and 2005, respectively. He previously worked as a Postdoctoral Scholar with the Department of Bioengineering, Stanford University, and as a full-time Lecturer with Kyungpook National University. He is currently an Associate Professor with the School of Computer Science and Engineering, Chung-Ang University. His research interests include computer graphics, computational bioengineering, multi-scale visualization, geometric modeling, and image processing for biomedical applications.



JAESUNG LEE received the B.S., M.S., and Ph.D. degrees in computer science from Chung-Ang University, Seoul, Republic of Korea, in 2007, 2009, and 2013, respectively. He is currently the Head and an Associate Professor with the Department of Artificial Intelligence, Chung-Ang University. Concurrently, he serves as the Chief of the AI/ML Innovation Research Center, Chung-Ang University. His research focuses on classification and feature selection, particularly multilabel learning with information theory. His research interests include machine learning, multilabel learning, model selection, and neural architecture search.

...