IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 November 2023, accepted 8 December 2023, date of publication 25 December 2023,
date of current version 12 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3346408

RESEARCH ARTICLE

Sketch-Guided Latent Diffusion Model for High-Fidelity Face Image Synthesis

YICHEN PENG[1], CHUNQI ZHAO[2], HAORAN XIE[1], (Member, IEEE), TSUKASA FUKUSATO[3], AND KAZUNORI MIYATA[1], (Member, IEEE)

[1]Japan Advanced Institute of Science and Technology, Nomi-shi, Ishikawa 923-1292, Japan
[2]School of Creative Informatics, The University of Tokyo, Bunkyo, Tokyo 113-8654, Japan
[3]School of Fundamental Science and Engineering, Waseda University, Shinjuku, Tokyo 169-8555, Japan

Corresponding author: Yichen Peng (yichen.peng@jaist.ac.jp)

This work was supported in part by JST SPRING under Grant JPMJSP2102, in part by the Kayamori Foundation of Informational Science Advancement, and in part by the Waseda University Grant for Special Research Project 2023C-436.

ABSTRACT Synthesizing facial images from monochromatic sketches is one of the most fundamental tasks in the field of image-to-image translation. However, it is still challenging to teach model high-dimensional face features, such as geometry and color, and to the characteristics of input sketches, which should be considered simultaneously. Existing methods often use sketches as indirect inputs (or as auxiliary inputs) to guide models, resulting in the loss of sketch features or in alterations to geometry information. In this paper, we introduce a Sketch-Guided Latent Diffusion Model (SGLDM), an LDM-based network architecture trained on the paired sketch-face dataset. We apply a Multi-Auto-Encoder (AE) to encode the different input sketches from the various regions of a face from the pixel space into a feature map in the latent space, enabling us to reduce the dimensions of the sketch input while preserving the geometry-related information of the local face details. We build a sketch-face paired dataset based on an existing method XDoG and Sketch Simplification that extracts the edge map from an image. We then introduce a Stochastic Region Abstraction (SRA), an approach to augmenting our dataset to improve the robustness of the SGLDM to handle arbitrarily abstract sketch inputs. The evaluation study shows that the SGLDM can synthesize high-quality face images with different expressions, facial accessories, and hairstyles from various sketches having different abstraction levels, and the code and model have been released on the project page. https://puckikk1202.github.io/difffacesketch2023/

INDEX TERMS Diffusion model, image synthesis, sketch-guided image generation.

I. INTRODUCTION

Synthesizing images, especially human faces, from a monochromatic sketch is one of the most fundamental tasks in the image-to-image translation, as it benefits various applications, such as character design and inmate tracking. However, the sparse distributions of single-channel sketch data challenge feature extraction and generalization. In addition, collecting paired datasets of painter's sketches and the corresponding photographs is time-consuming and labor-intensive; similarly, it is challenging for the synthesis model to understand the monochromatic sketch input with redundant semantic information (e.g., separated facial components, expressions, accessories, and hairstyles).

Generative Adversarial Network-based-based generative models [10], [16] are one of the most feasible solutions for sketch-to-image generation based on semantic mask-annotated datasets [2], [10]. Despite allowing users to arrange facial semantics (i.e., regional-only conditions), many details may be lost or arbitrarily synthesized, such as wrinkles and mustaches. Instead of applying semantic masks, previous GAN-based models [3] trained using sketch-face paired datasets can directly generate (and edit) face images from monochrome sketches. However, they are unsuitable for

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

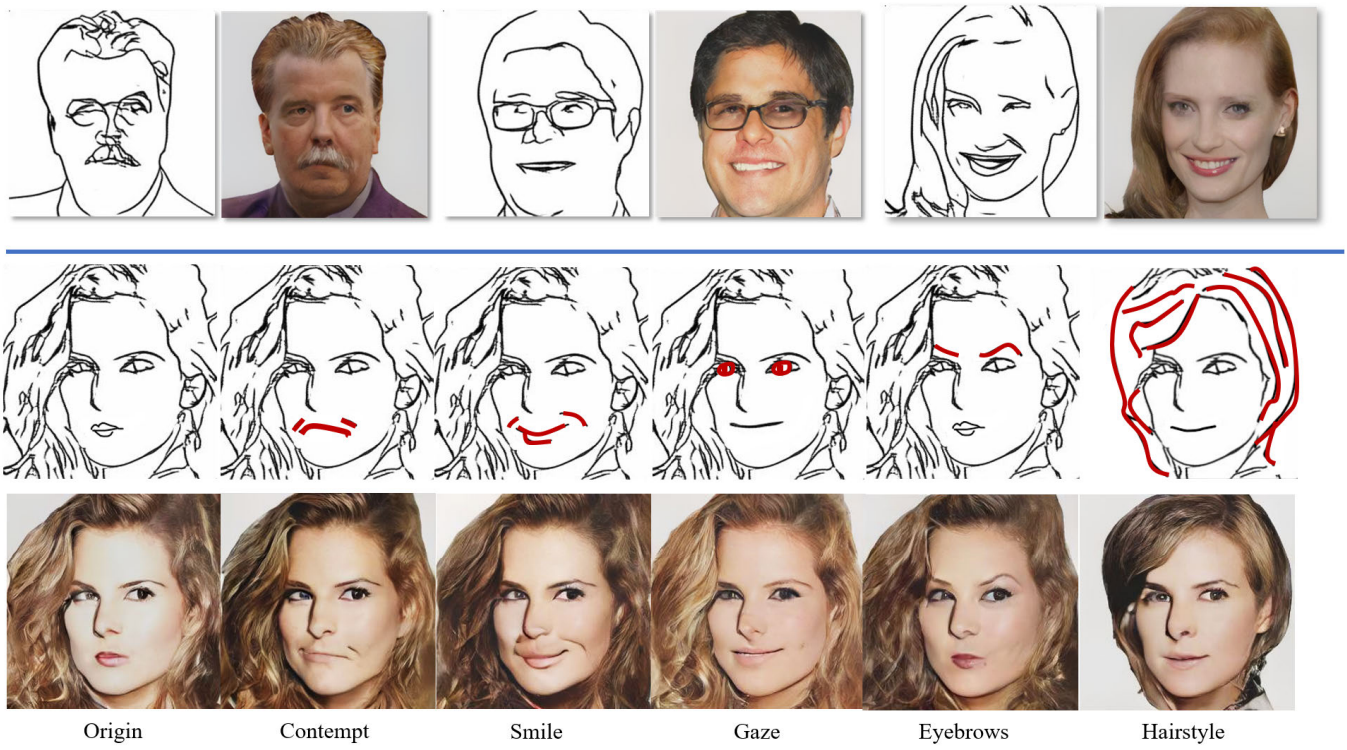

FIGURE 1. A Sketch-Guided Lantent Diffusion Model (SGLDM) synthesizes high-quality face images with a high consistency with input sketches, SGLDM enables users to simply edit face images, such as different expressions, facial components, hairstyle, etc. The edited strokes are highlighted in red.

handling local geometrical details, such as accessories and expressions, as no semantic information was directly specified in the rough monochromatic sketches. More recently, the diffusion model (DM) [8], [14], [25] and Contrastive Language-Image Pre-training (CLIP) [18] have achieved tremendous success with the text-to-image task. However, in the case of image-to-image transformation, especially sketch-to-image, their system requires not only an image input but also appropriate text inputs, and it may not generate desired images, as shown in Figure 2. The other conditioning-guided DM-based models, such as ILVR [4] and SDEdit [12], approached the image-to-image task by inputting an RGB image reference to control the synthesis. However, it is generally difficult to specify image details after noise injection and resampling of the query input.

To maximize the generative models to learn from the paired sketch to gather more accurate information, in this work, we introduce a Sketch-Guided Latent Diffusion Model (SGLDM), a network architect trained using a sketch-face dataset. The LDM is exceptional at flexible and high-quality inference under different conditions, so we apply an LDM as a backbone for our sketch-guided image synthesis training. We also apply a Multi-Auto-Encoder (AE) to encode query sketches from the pixel space into feature maps in the latent space of the image feature, enabling us to reduce the dimensions of the sketch input while preserving the geometrical-related information of the face's local details. Moreover, we apply a two-stage training process to achieve better distribution mapping between the sketch and image domains. Because different people focus on different facial regions, this often leads to varied levels of abstraction in the input sketch. For example, some people focus on the details of the eyes, while others focus on the mouth. To access sketch data with different levels of abstraction, we introduce a data-augmentation method, named *Stochastic Region Abstraction* (SRA) to improve the robustness of SGLDM, while sketch data are extracted from Celeba-HQ using sketch simplification methods [23], [24], [29]. The evaluation study shows our model can generate natural-looking face images from sketches with different levels of detail. In addition, the SGLDM also enables users to synthesize desired face images (at a resolution of 256 $\times$ 256) with different expressions, facial accessories, and hairstyles via a monochromatic sketch (see Figure 1).

In short, our main contributions are summarized as follows.

- We proposed SGLDM, a sketch-input-only model, trained via a two-stage training process to synthesize faces with high quality and input consistency.
- We introduced SRA, a data augmentation strategy for synthesizing convincible faces from input sketches at different levels of abstraction.
- We verified the SGLDM achieves superior scores in various metrics compared to state-of-the-art methods and it is sufficiently robust to generate the intended face images.

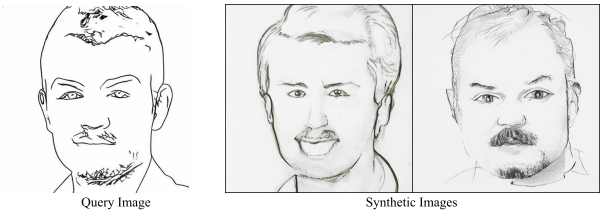

FIGURE 2. An example of implementing an LDM-based model, *stable diffusion* [20] with pre-trained weights. Although we inputted a single sketch (left) and texts (e.g., "*a face photo*" or "*a portrait*"), the generated results are not colored images but monochromatic sketches, and they do not reproduce the contours of the input sketch.

II. RELATED WORK

A. SKETCH-BASED IMAGE SYNTHESIS

The field of image synthesis from sketches has witnessed significant exploration over the past decade. From its inception, the problem of sketch-to-image translation has been tackled as an image-to-image transformation task. As such, researchers have endeavored to train deep learning-based networks to bridge the gap between monochromatic sketches and full-color RGB images.

Several supervised GAN-based generative models, such as Pix2Pix [36] and Pix2pixHD [27], rely on paired sketch-image datasets, which are created by extracting the edge information from real images to facilitate model training.

To enhance the efficacy of translating the image domain into the sketch domain, the construction of a substantial corpus of paired sketches and photographs is imperative. Consequently, such datasets as Sketchycoco [6], which categorize objects into distinct classes, have been introduced. However, concerning to facial sketch image datasets, the availability is notably limited, as exemplified by the datasets CUHK Face Sketches [28], [35].

Conversely, unsupervised image-to-image translation methods, such as CycleGAN [36] and DualGAN [32], have been explored in other works. More recently, with the burgeoning advancement in disentangled representation within StyleGAN's $w+$ space, sketch-to-sketch translation has been treated as a style transfer task, illustrated by methods like DualStyleGAN [30] and Pixel2Style2Pixel [19]. Furthermore, akin to the recently popular text-to-image generative models, GAN-based approaches also permit users to generate and edit images via textual and sketch inputs, such as [15] and [17].

Nonetheless, end-to-end GAN-based models have been associated with such issues as unstable training and susceptibility to overfitting on specific datasets. These challenges restrict the diversity and quality of synthesized results. Hence, drawing inspiration from the notable performance of LDMs in conditional image synthesis tasks, we propose the SGLDM as a solution for achieving high-quality face synthesis with enhanced input consistency.

B. DIFFUSION MODELS

In recent times, diffusion and score-based models have emerged as formidable contenders in the realm of image synthesis, a fundamental component of which is the U-Net architecture [21], lauded for its excellence in fostering diversity, ensuring quality, stabilizing training, and offering module extensibility.

Noteworthy advancements have been presented in previous studies, such as [8] and [25], demonstrating a superior performance, particularly in the domain of unconditional image synthesis. However, a significant impediment remains the hefty computational costs, which subsequently constrain the resolution of the images produced. In a serendipitous turn of events, contributions by [20] have provided a solution. In their approach, images are initially encoded from a high-dimensional RGB space to a more manageable, low-dimensional latent feature space.

Subsequently, this latent representation is employed to navigate both the forward and backward diffusion processes. In addition, the architecture's commendable modular extensibility equips it to handle a plethora of image-to-image tasks. This includes, but is not limited to, image inpainting, semantic mask-to-image translation, and layout-to-image generation, as elucidated by [20].

While certain methodologies have focused on altering the network architecture of DM-based designs, alternative approaches start from ILVR [4] and SDEdit [12]. Upto more recently represented by ControlNet [13] and T2I-Adatper [34], many have chosen a different path. However, these strategies involve fine-tuning the models with supplementary plug-in condition modules. Alternatively, during the sampling process, they incorporate extra constraint loss functions to govern the sampling procedure. This concerted effort has yielded impressive results, enabling models to excel in the task of generating high-quality images from sketches. Intriguingly, their method necessitates a blurry RGB reference, which serves the dual purpose of iteratively guiding the sampling process and acting as an input reference. Despite their efforts, the delineation of intricate image details was compromised, primarily attributable to the conditioning's inherent blurriness. Meanwhile, when considering the sketch-to-image task, a paramount challenge emerges: the monochromatic sketches inherently possess a dearth of semantic information. Consequently, executing the sketch-to-image transformation via DMs invariably demands supplementary inputs, such as auxiliary text prompts, to compensate for this information void.

III. METHOD

A. OVERVIEW

Our goal is to synthesize a high-quality face images that are highly consistent with the input sketch. We assume that the feature distribution of the monochromatic sketches in the dataset is much more irregular and sparser than that of the RGB images. Jointly training a model to encode the sketch embedding, and map the sketch domain to the image domain may result in a rough distribution (see Figure 4 [dashed line])). Therefore, we implemented a two-stage training

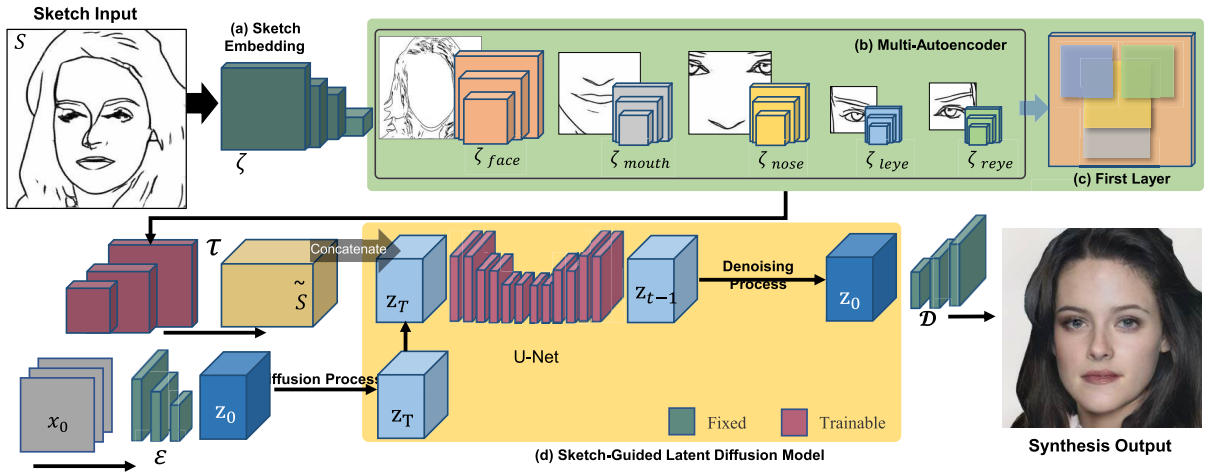

FIGURE 3. The framework of SGLDM. In the sketch-embedding stage (a), given a sketch input S, a pretrained sketch encoder ζ which consists of $\zeta = \{\zeta_{\text{leftEye}}, \zeta_{\text{rightEye}}, \zeta_{\text{nose}}, \zeta_{\text{mouth}}, \zeta_{\text{face}}\}$ encodes S into a feature vector $S'_{\text{leftEye}}, S'_{\text{rightEye}}, S'_{\text{nose}}, S'_{\text{mouth}}, S'_{\text{face}}$. We then combine the feature into an overall feature map (c). The decoder θ then decoders (c) into a feature map $\tilde{S}$. In the latent denoising stage (b), the random latent code Z_T is concatenated by the feature map $\tilde{S}$ and denoised to Z_0 by a U-Net. Finally, the output latent code Z_0 is decoded by $\mathcal{D}$ to the final output.

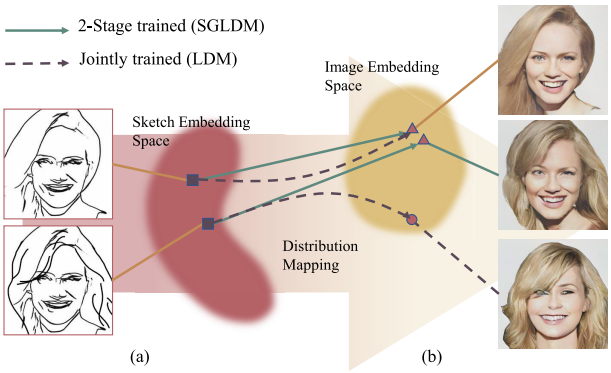

FIGURE 4. The illustration of feature distribution mapping between the jointly-trained conditional embedding (dashed line) and the separately-trained conditional embedding (solid line), from the sketch (a) to the image (b) domain.

method to optimize the distribution mapping between the sketch and image domains, as shown in Figure 4 (solid line). More details can be found in Section IV.

B. PRELIMINARIES

DMs represent a subset of generative models, underpinned by two fundamental mechanisms: first, the diffusion process, often referred to as the forward process, incrementally introduces Gaussian noise to the dataset through a predetermined Markov chain spanning T steps. Second, the denoising process, an algorithmic model trainable to synthesize samples deriving from Gaussian noise. Furthermore, DMs are versatile, with the ability to be contingent on other inputs. For instance, in the context of text-to-image DMs, text can be used as an input condition. Central to the operational efficacy of these models is their training objective. Typically, for a DM, this objective is symbolized as ϵ_θ, which essentially predicts the noise, and it objective is typically framed as a streamlined iteration of the variational bound:

$$L_{\text{DM}} = \mathbb{E}_{x_0,\epsilon\sim N(0,I),c,t}[\|\epsilon_t - \epsilon_\theta(x_t, c, t)\|^2] \tag{1}$$

where x_0 denotes the authentic data, which are augmented with a conditioning element, represented as c. The diffusion process temporal progression is captured by t, which ranges within $[0, T]$. This acts as a chronological gauge, signaling the evolution of the diffusion process across its steps. The noisy data at a particular time step t are symbolized by x_t, a function of the genuine data $x0$, the Gaussian noise ϵ, and the predefined coefficients α_t and σ_t. Specifically, $x_t = \alpha_t x_0 + \sigma_t \epsilon$ articulates the amalgamation of the real data and noise, moderated by α_t and σ_t, the roles of which are pivotal. They are not arbitrary but are systematically defined functions of t, and their values influence the trajectory and intensity of the diffusion process. After successfully training the model ϵ_θ, it is then empowered to generate visual content, or images, starting from arbitrary noise. This generation is not instantaneous, but unfolds iteratively, resembling the incremental character of the diffusion process.

In a more recent development, a method called LDM, as introduced by Rombach and colleagues [20], has emerged with the aim of mitigating computational costs. The rationale behind this approach hinges on the observation that, even after passing through the neural network of the AE model, certain features that contribute to perceptual intricacies and semantic significance remain embedded within the latent code.

The LDM technique involves the utilization of a pre-trained encoder denoted as ε, which is tasked with

encoding an image x residing in a high-dimensional RGB space, represented as $x \in \mathbb{R}^{H\times W\times 3}$, into a lower-dimensional latent code z, situated in $z = \varepsilon(x) \in \mathbb{R}^{h\times w\times 3}$. Subsequently, a pre-trained decoder D is employed to reverse this process, effectively generating images from the latent code, denoted as $\tilde{x} = \mathcal{D}(z)$. The significance of this transformation lies in its potential to facilitate a transition in the loss-term L_{LDM}.

$$L_{\text{LDM}} = \mathbb{E}_{x_0,\epsilon\sim N(0,I),c,t}[\|\epsilon_t - \epsilon_\theta(z_t, c, t)\|^2] \quad (2)$$

IV. SKETCH-GUIDED LATENT DIFFUSION MODEL

A. FRAMEWORK

In the context of face synthesis, our objective is to generate facial images based on a single provided sketch input. To achieve this, we treat the sketch as a guiding condition for the model during the denoising process. The framework of our approach, denoted as SGLDM, is illustrated in Figure 3(a, d). Drawing inspiration from the work of Rombach and colleagues [20], we also incorporate an LDM to optimize computational efficiency.

Incorporating our sketch-condition pairs, the training loss L_{SGLDM} for the conditional LDM can be expressed as follows:

$$L_{\text{SGLDM}} = \mathbb{E}_{x_0,\epsilon\sim N(0,I),c,t}[\|\epsilon_t - \epsilon_\theta(z_t, \tau_\theta(\tilde{S}), t)\|^2] \quad (3)$$

In this formulation, $\tilde{S}$ represents a sketch feature that has been encoded by a pre-trained sketch encoder, denoted as $\zeta(S)$, operating on the input sketch S. The τ_θ function serves as a decoder, responsible for estimating a conditional map allows for the reversal of the diffusion process applied to $\varepsilon(x)$. It's important to note that both τ_θ and ϵ_θ are concurrently trained.

Instead of solely training a sketch encoder to generate a conditional feature map for Z_T to facilitate denoising, we introduce a "Conditioning Module" by pretraining a *Multi-AE* network architecture. Drawing inspiration from previous works that segmented the global facial structure into local components for individual networks, as seen in DeepFaceDrawing [3], APDrawGAN [31], and MangaGAN [26], our overarching encoder ζ comprises five distinct partial encoders, denoted as $\zeta = \{\zeta_{\text{leftEye}}, \zeta_{\text{rightEye}}, \zeta_{\text{nose}}, \zeta_{\text{mouth}}, \zeta_{\text{face}}\}$, as illustrated in Figure 3(b,c).

For face editing, as opposed to face synthesis, we introduce an original facial input. To facilitate this, we train a VQVAE, which employs vector quantization to enhance the quality of image synthesis by learning discrete latent representations, using our sketch dataset to encode the dilated sketch. Both the original face and the sketch input are inversely masked, with the face being masked by the region designated for editing and the sketch input being masked by the remaining area. Subsequently, we concatenate the two encoded features with an additional binary mask map to train the LDM.

B. 2-STAGE TRAINING STRATEGY

Our training phase incorporates a two-stage process. During the sketch embedding phase, we pre-train the *Conditioning Module*. This pre-training is achieved by minimizing the cumulative Mean Squared Error (MSE) loss $L_{\text{Multi-AE}}$, stemming from each individual partial encoder. This can be mathematically formulated as:

$$L_{\text{Multi-AE}} = \| \sum_{\zeta_i\in\zeta_{\text{AE}}} \zeta_i(x) - x\|_2 \quad (4)$$

where $\zeta_{\text{AutoEncoder}}$ denotes the *Multi-AE*, where ζ_i represents the autoencoder for each facial region i. This includes an encoder ζ_{leftEye} and a corresponding decoder ζ'_{leftEye}, among others. It should be noted that these decoders are exclusively employed during the training phase of the *Multi-AE* and are not utilized in either the training stage of the conditional LDM or the overall inference stage.

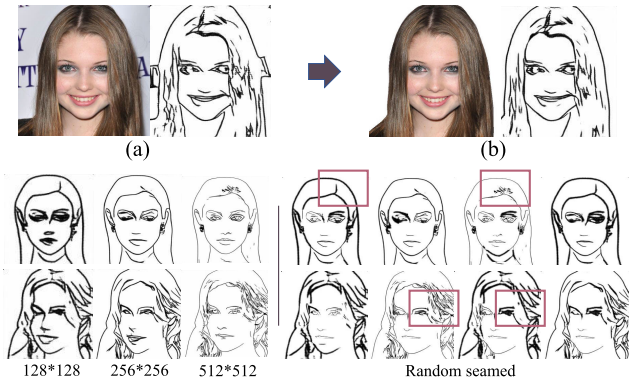

FIGURE 5. The original image in Celeba-HQ and its extracted edge map (a), and the result of paired data after removing the background (b). Sketch simplification results from 3 different resolutions faces (bottom-left), and the random seamed data samples (bottom-right).

Our decision to initiate pre-training with the *Multi-AE* rather than proceeding directly to joint training for the sketch encoders-thereby providing a conditional feature map for the SGLDM-is underpinned by two core motivations:

- **Domain Distribution Alignment:** Our intent is to cultivate a model that more adeptly discerns and maps the relationships between the distinct domain data distributions characterizing the sketches and faces. In doing so, we aim to yield a seamlessly integrated domain distribution space, as visualized in Figure 4.
- **Computational Efficiency:** Adopting a two-stage training strategy provides computational advantages. Specifically, by decoupling the trainable parameters of models across distinct stages, we streamline and enhance the model optimization process.

In the subsequent training phase, and to bolster the SGLDM's adaptability across a variety of sketches, we incorporate what we term the *arbitrarily masking conditional training strategy*, inspired by the *Masked AutoeEncoder* [7], involves randomly occluding segments of the input, leaving it to the model to these sections. Given the pre-trained status of our sketch encoder, ζ, our approach specifically entails the random masking of the conditioning feature map, S, during the training phase for denoising U-Net.

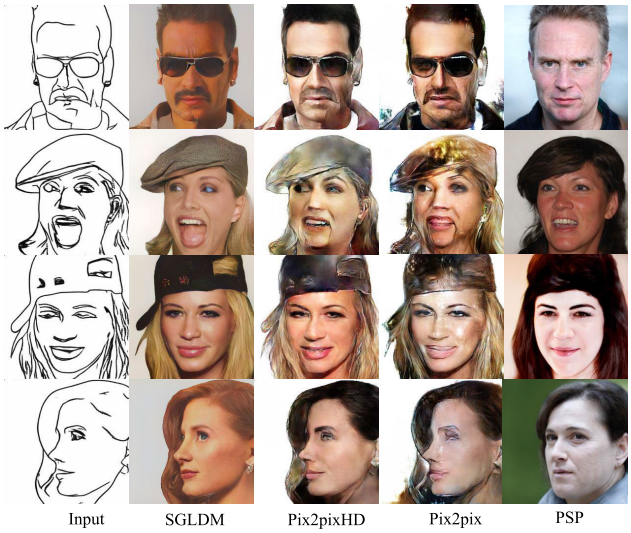

FIGURE 6. Examples of corner cases of sketch inputs that featured glasses, a hat, or a side profile.

C. STOCHASTIC REGION ABSTRACTION DATA AUGMENTATION

To build our training dataset, we use 10,000 high-quality face images from the Celeba-HQ dataset [10]. We first remove the background of the photos, as shown in Figure 5(a,b). Next, we utilize *sketch simplification* [23], [24] to generate edge maps of the faces. To enhance the robustness of the SGLDM to manage sketch inputs with arbitrary abstraction, we introduce SRA to augment the dataset. We observed that the abstraction levels of the extracted edge maps depend on the image resolution. As such, we resized the original photos into 128×128-, 256×256-, and 512×512- pixel resolutions respectively (see Figure 5 [left-bottom]), and we augmented our sketch dataset. The red box highlighted a clear difference in the different abstraction levels in the hair and eye regions. Moreover, following our *Multi-AE* related region of every single encoder, we crop the edge maps into five different pieces and randomly combine them back together to form a new edge map with random seams at different abstraction levels, as shown in Figure 5 (bottom-right). We finally utilized 8,000 images for training, 1,000 for validating, and 1,000 for testing.

V. EXPERIMENT AND RESULTS

We conducted several experiments to verify the quality and sketch input consistency of the SGLDM's synthetic face images.

A. IMPLEMENTATION

Both stages of SGLDM are trained on a single NVIDIA RTX3090 GPU. In stage one *Multi-AE* training, the training is performed for 500 epochs with an Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a batch size 64. The dimensions of the latent space of every AE are the same at 512. In stage two, our SGLDM is trained on 300 epochs with an Adam optimizer as well, but the batch size is 8. The feature map of the sketch embedding has 8 channels, plus three channels of the LDM latent size, our denoising U-Net input is 11 channels of latent code and the output is 3 channels.[1]

TABLE 1. Preference results of user study.

Method	Quality	Fidelity
Pix2pixHD [27]	17%	27%
Psp [19]	37%	2%
Ours (SGLDM)	**46%**	**71%**

B. QUANTITATIVE COMPARISONS

We compare SGLDM with several state-of-the-art image-to-image translation methods on the sketch2face task (pip2pixHD [27], pix2pix [36], DeepFaceDrawing [3], pixel2style2pixel (PSP) [19], and Palette [22]). We re-trained most of the models on our 10K faces dataset picked from Celeba-HQ in the same training settings. We directly implement the pre-trained weight based on 512×512- of the DeepFaceDrawing Model.

For the overall quality of results from different competing methods, as shown in Figure 11, the SGLDM synthesizes more realistic faces while more faith is placed in the input sketch. Pix2pix, Pix2pixHD, and DeepFaceDrawing, however, tended to synthesize noisy faces when faces' sketches were not facing straight, such as the third and the last columns. Note that DeepFaceDrawing additionally required a condition to control the gender of synthetic faces, so we prepared both of the results. Although PSP achieved higher quality results visually than other methods, their methods showed poor fidelity of the sketch. Besides, Palette, one DM-based image-to-image translation method, failed to synthesize convincible faces from a sketch-only input. To our knowledge, there is less state-of-the-art DM-based trained from-scratch pipeline that relies only on monochromatic sketch input, (In addition to some finetune on pre-trained models, such as ControlNet [34] and T2I-adapter [13]) and most are based on text2image, segmap2image, or image inpainting pipelines fused with sketch input (e.g., [9]).

Next, we compared SGLDM, Pix2pixHD, and PSP which have similar fidelity results (see Figure 12). The black strokes on the right are input sketches for synthesizing the left-face images, and the red strokes behind the black strokes are filtered versions of the synthesized images using Adobe Photoshop's *sketch filter tool* [1].

From the results, we confirm that the SGLDM can synthesize noiseless faces maintaining maximum consistency with the input sketch, except for some facial details, such as nasolabial folds. Figure 6 shows examples of generated face images with expressions, accessories, and hairstyles,

[1]Check up the codes and pre-trained model in our project page.

TABLE 2. Quantitative comparisons. We applied the Fréchet inception distance (FID) (↓) score to measure the synthetic faces quality, the Learned Perceptual Image Patch Similarity (LPIPS) (↓) scores to evaluate the consistency between real faces and synthesized results, and a recall ratio (REC↑) to evaluate the input consistency.

	Low abstraction			Mid abstraction			High abstraction		
Method	FID↓	LPIPS↓	REC↑	FID↓	LPIPS↓	REC↑	FID↓	LPIPS↓	REC↑
Pix2pix [36]	53.67	0.20	0.54	59.46	0.23	0.50	63.45	0.28	0.51
Pix2pixHD [27]	51.23	0.18	0.62	53.71	**0.22**	0.55	60.23	0.25	0.53
Psp [19]	83.48	0.29	0.37	83.32	0.26	0.45	85.54	0.28	0.48
SGLDM *joint training*	46.28	0.20	0.65	48.62	0.23	0.54	50.33	0.26	0.51
SGLDM w/o *SRA*	**38.57**	**0.17**	**0.77**	48.87	0.26	0.51	57.76	0.29	0.48
Ours (SGLDM)	43.58	0.22	0.71	**45.46**	0.24	**0.59**	**46.83**	**0.24**	**0.57**

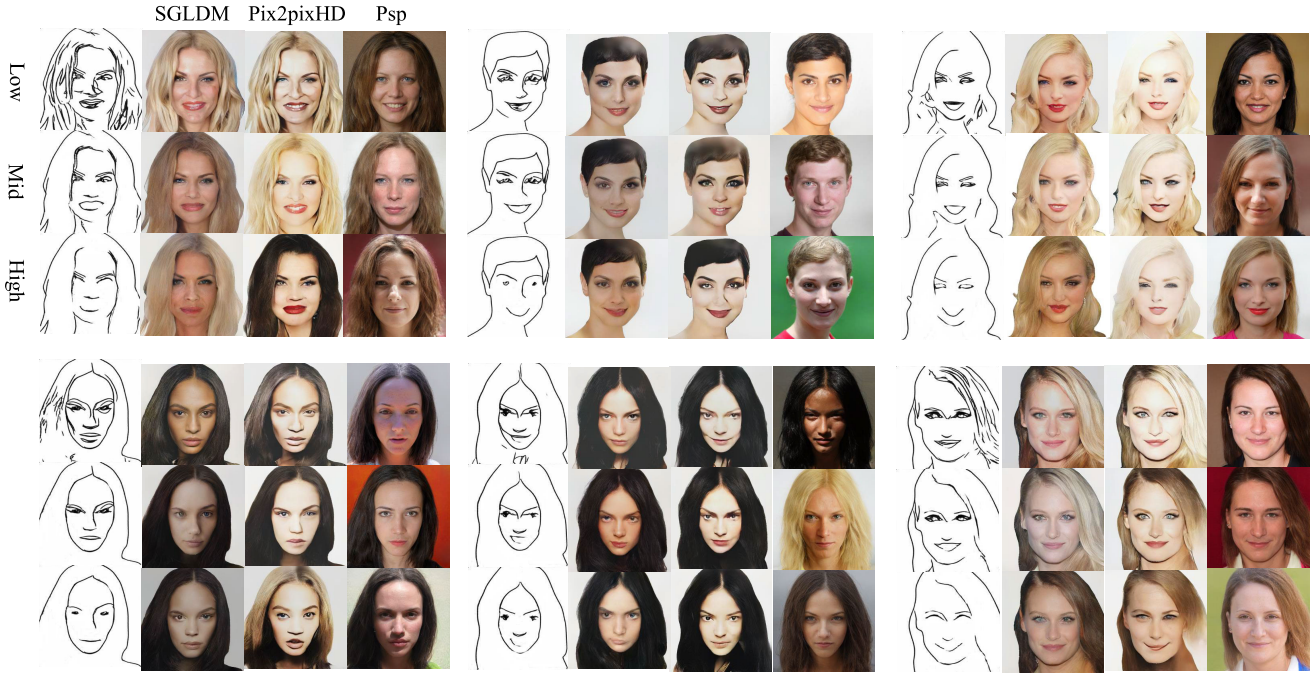

FIGURE 7. The comparison of synthetic results of different sketch inputs having three abstraction levels.

demonstrating that our method achieves a better balance between visual quality and input consistency of inputs.

We further conducted a user study to compare the visual quality and the input consistency of three methods: SGLDM, Pix2pixHD, and PSP, some of the compared samples are shown in Figure 6. Note that Pix2pix was not included as its visual quality is similar to that of Pix2pixHD. Participants were asked to choose their preferences among the three types of synthetic face images generated from different models for both visual quality and input consistency. From Table 1, we confirmed that face images synthesized by the SGLDM achieved the highest preference for input consistency and visual quality, not dissimilar to PSP.

C. QUALITATIVE EVALUATION

Concerning the input consistency, we calculated the recall ratio (REC) between the black and red strokes (see Figure 12). In addition, as the visual differences in the output are minimal with different resolutions of input sketches (as mentioned in Section IV-C), we prepared input sketches by manually erasing some strokes from the original sketches, and we generated face images (see Figure 7). From these results, we confirmed that SGLDM is robust enough to handle rough sketches with different abstraction levels.

We also conducted an ablation study to compare the metrics scores between the joint training & two-stage training methods and to verify the validity of our SRA strategy, as shown in Table 2 (lower-rows). We observed that the scores of SGLDM trained via joint training methods showed a similar performance to Pix2pixHD. Although SGLDM instructed on a single abstraction level dataset (without SRA), shows the best performance on highly detailed sketch input (low abstraction), it declined significantly under higher abstraction inputs, as shown in Figure 8 (rightmost column,

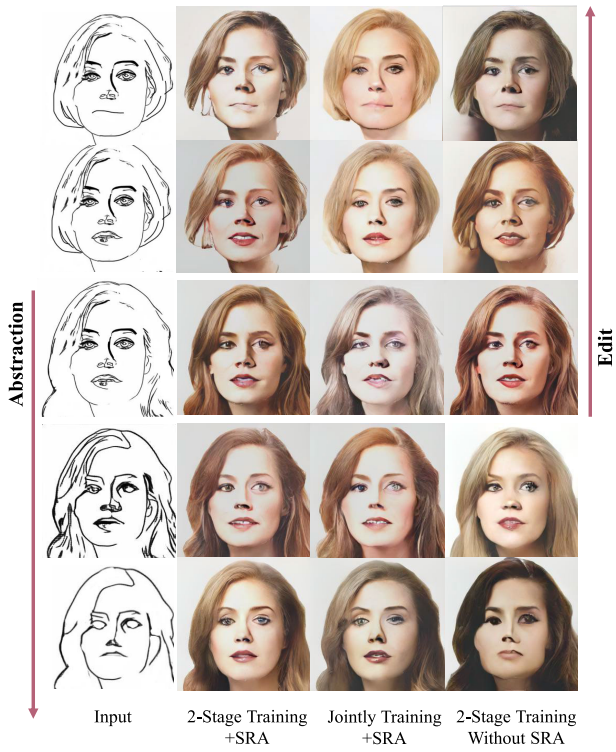

FIGURE 8. The synthetic faces of the ablation study.

downward). From the results, both separate training and data augmentation methods improved the overall performance of the SGLDM on various sketch inputs (see Figure 8 (second column)).

D. EDITING CAPABILITY

We considered the usefulness of face editing. Figure 9 shows examples of partial editing, (a,b) hair styles of both males and females, (c) the earings, and (d) expressions. In addition, we compared the synthetic faces of the two-stage trained model and the jointly trained model (see Figure 8)., illustrating that the two-stage trained SGLDM is more robust than the jointly trained SGLDM,forming a different identity easily after editing, (see Figure 8 [third column, upward]). As a result, the SGLDM is sufficiently robust enough to edit the intended face at will using the synthetic results.

VI. LIMITATIONS AND FUTURE WORK

Although the SGLDM achieves high consistency with input sketches, the synthesized result tends to be too strongly affected by the input sketches. That is, noise and artifacts might be generated when inputting extremely poor sketches, as shown in Figure 10. To solve this issue, some trade-off methods or algorithms will be required to keep the balance between inputs' consistency and outputs' convincibility. In addition to monochromatic sketch input, we plan to consider a method of inputting several color cues to handle color information such as skin and hair regions. Moreover, we evaluated SGLDM's performance on the face

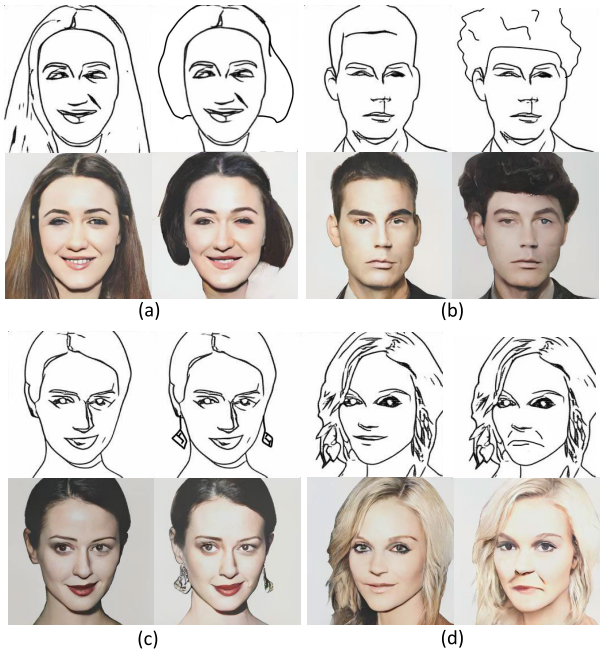

FIGURE 9. Examples of face editing with SGLDM. (a,b) hairstyles, (c) earrings, and (d) expressions.

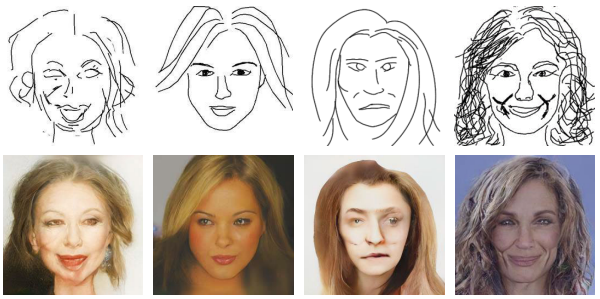

FIGURE 10. Less successful examples generated from the low-quality sketch inputs. Except for the second sketch from the right, the input sketches are from [3].

synthesizing task. We believe that a similar framework can also be applied into other sketch-image tasks by changing the training dataset such as, Large-scale Scene Understanding (LSUN) [33] and (Animal Face HQ) AFHQ [5]. In addition SRA can simply augmented each dataset to enhance the robustness of each model.

In this paper, although we implemented an LDM-based method to reduce the computation costs, the SGLDM (i.e., the training and the sampling stage) is still computationally heavier than GAN-based models. In the training stage of a 256 × 256 model, the maximum batch size on a single NVIDIA RTX3090 is 8, while a 512×512 model's maximum batch size is only 1, and in the sampling stage, the average time cost of one image is around 15.2 seconds. Although it can be cut down to 50 sampling steps, and takes around 5 to 6 seconds when using the denoising diffusion implicit models (DDIM) sampling strategy, the current implementation is still difficult to incorporate into a real-time interactive graphical

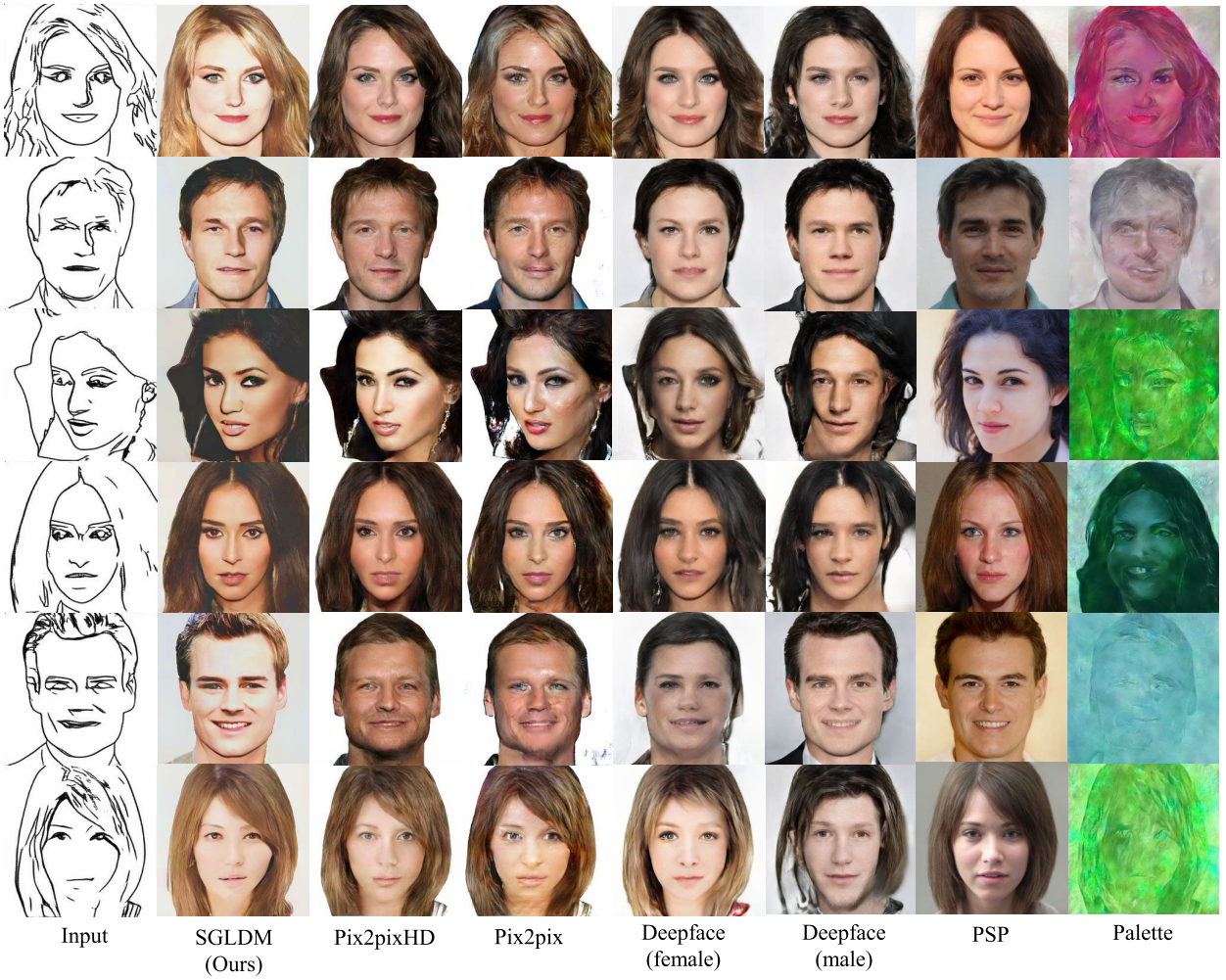

FIGURE 11. Qualitative comparisons of the proposed SGLDM with the state-of-the-art methods.

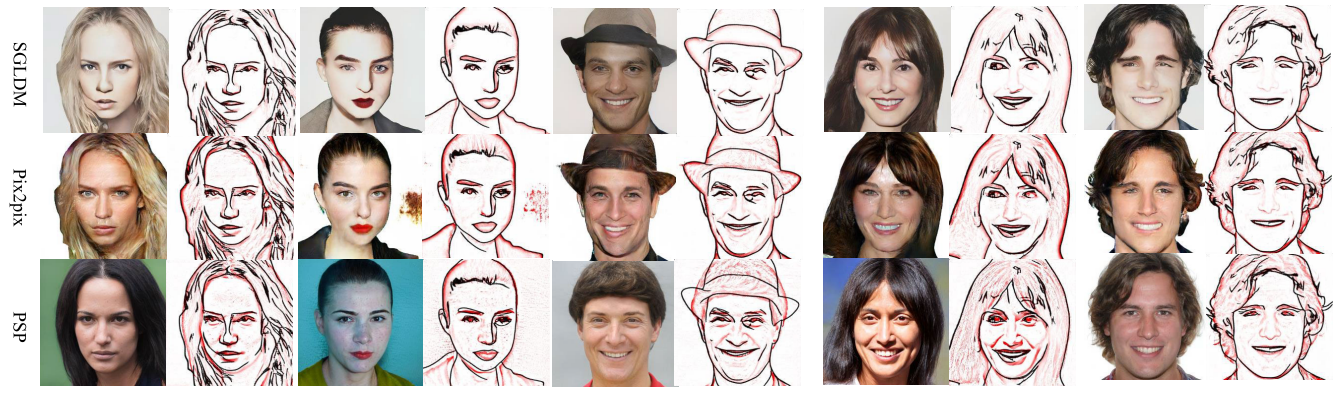

FIGURE 12. Fidelity comparisons of the proposed SGLDM with competing methods.

user interface (GUI). Furthermore, the latest methods like the Latent Consistency Models (LCMs) [11] significantly reduce the sampling generation time by requiring only 2 to 4 steps for sampling. We plan to reference their approach for application in our SGLDM to enable more real-time interaction in the future.

VII. CONCLUSION

This paper has proposed SGLDM, an LDM-based architect face synthesizing model with a Multi-AE to encode the query sketch as a conditional map while preserving the geometrical-related information of the face's local details. We also introduced SRA, a data-augmentation strategy that

enables the models to deal with a sketch input of different abstraction levels. We conducted experiments to verify that the SGLDM could synthesize high-quality face images with high input consistency. Moreover, the SGLDM is robust enough to edit the synthetic results with different expressions, facial accessories, and hairstyles.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the editor for their insightful comments, which improved this manuscript.

REFERENCES

[1] Adobe Systems. (2022). *Photo to Pencil Sketch*. [Online]. Available: https://www.adobe.com/creativecloud/photography/discover/photo-to-pencil-sketch.html

[2] H. Caesar, J. Uijlings, and V. Ferrari, "COCO-stuff: Thing and stuff classes in context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1209–1218.

[3] S.-Y. Chen, W. Su, L. Gao, S. Xia, and H. Fu, "DeepFaceDrawing: Deep generation of face images from sketches," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 1–16, Aug. 2020.

[4] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "ILVR: Conditioning method for denoising diffusion probabilistic models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 14347–14356.

[5] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 8185–8194.

[6] C. Gao, Q. Liu, Q. Xu, L. Wang, J. Liu, and C. Zou, "SketchyCOCO: Image generation from freehand scene sketches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 5173–5182.

[7] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 15979–15988.

[8] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[9] D. Horita, J. Yang, D. Chen, Y. Koyama, K. Aizawa, and N. Sebe, "A structure-guided diffusion model for large-hole image completion," 2022, *arXiv:2211.10437*.

[10] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, Jun. 2020, pp. 5548–5557, doi: 10.1109/CVPR42600.2020.00559.

[11] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent consistency models: Synthesizing high-resolution images with few-step inference," 2023, *arXiv:2310.04378*.

[12] C. Meng, Y. Song, J. Song, J. Wu, J. Zhu, and S. Ermon, "SDEdit: Image synthesis and editing with stochastic differential equations," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–33.

[13] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2I-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," 2023, *arXiv:2302.08453*.

[14] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, Jul. 2021, pp. 8162–8171. [Online]. Available: https://proceedings.mlr.press/v139/nichol21a.html

[15] U. Osahor and N. M. Nasrabadi, "Text-guided sketch-to-photo image synthesis," *IEEE Access*, vol. 10, pp. 98278–98289, 2022.

[16] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2332–2341.

[17] M. Pernu, C. Fookes, V. Truc, and S. Dobriek, "FICE: Text-conditioned fashion image editing with guided GAN inversion," 2023, *arXiv:2301.02110*.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763. [Online]. Available: http://proceedings.mlr.press/v139/radford21a/radford21a.pdf

[19] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: A styleGAN encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 2287–2296.

[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 10674–10685.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[22] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *Proc. ACM SIGGRAPH Conf.*, New York, NY, USA, Aug. 2022, p. 15.

[23] E. Simo-Serra, S. Iizuka, and H. Ishikawa, "Mastering sketching: Adversarial augmentation for structured prediction," *ACM Trans. Graph.*, vol. 37, no. 1, pp. 1–13, 2018.

[24] E. Simo-Serra, S. Iizuka, K. Sasaki, and H. Ishikawa, "Learning to simplify: Fully convolutional networks for rough sketch cleanup," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, 2016.

[25] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. 9th Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–20. [Online]. Available: https://openreview.net/forum?id=St1giarCHLP

[26] H. Su, J. Niu, X. Liu, Q. Li, J. Cui, and J. Wan, "MangaGAN: Unpaired photo-to-Manga translation based on the methodology of Manga drawing," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 3, pp. 2611–2619.

[27] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8798–8807.

[28] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Sep. 2018.

[29] H. Winnemöller, "XDoG: Advanced image stylization with eXtended difference-of-Gaussians," in *Proc. ACM SIGGRAPH/Eurographics Symp. Photorealistic Animation Rendering*. New York, NY, USA: Association for Computing Machinery, Aug. 2011, pp. 147–156, doi: 10.1145/2024676.2024700.

[30] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, "Pastiche master: Exemplar-based high-resolution portrait style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 7683–7692.

[31] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "APDrawingGAN: Generating artistic portrait drawings from face photos with hierarchical GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 10735–10744.

[32] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2868–2876.

[33] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao, "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop," 2015, *arXiv:1506.03365*.

[34] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Paris, France, 2023, pp. 3836–3847.

[35] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. CVPR*, Colorado Springs, CO, USA, Jun. 2011, pp. 513–520.

[36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2242–2251.

YICHEN PENG received the Graduate degree in animation design from the Guangdong University of Technology, in 2017, and the M.S. degree from the Japan Advanced Institute of Science and Technology, in 2021, where he is currently pursuing the Ph.D. degree. His research interests include human–computer interaction and computer graphics.

CHUNQI ZHAO received the B.E. degree from Fudan University, in 2017, and the M.E. degree from The University of Tokyo, in 2020, where he is currently pursuing the Ph.D. degree with the Creative Informatics Department. His research interests include human–computer interaction and representation learning. He mainly works on improving experience for programming and creativity.

HAORAN XIE (Member, IEEE) received the Ph.D. degree from the Japan Advanced Institute of Science and Technology (JAIST), in 2015. He was with the Computer Science Department, The University of Tokyo, from 2015 to 2018. He was also an Assistant Professor and a Senior Lecturer with JAIST, from 2018 to 2023, where he is currently an Associate Professor. His main research interests include interactive computer graphics and user interfaces. He is a member of ACM.

TSUKASA FUKUSATO received the Ph.D. degree from the Department of Pure and Applied Physics, Waseda University, in 2017. He was with the Graduate School of Information Science and Technology, The University of Tokyo, as an Assistant Professor (2017–2023). He has been a Lecturer with the School of Fundamental Science and Engineering, Waseda University, since 2023. His main research interests include computer graphics (CG) and human–computer interaction (HCI), such as cartoon design. He is a member of ACM.

KAZUNORI MIYATA (Member, IEEE) received the B.S. degree from Tohoku University, in 1984, and the M.S. and Ph.D. degrees from the Tokyo Institute of Technology, in 1986 and 1997, respectively. He has been a Professor with the Japan Advanced Institute of Science and Technology (JAIST), since 2002. Prior to joining JAIST, he was an Associate Professor with the Department of Imaging Art, Tokyo Institute of Polytechnics, from 1998 to 2002, and a Researcher with IBM Japan, from 1984 to 1998. His research interests include computer graphics, media art, and multimedia applications. He is a member of ACM.