One-shot Face Sketch Synthesis in the Wild via Generative Diffusion Prior and Instruction Tuning

HAN WU, Guangdong University of Technology, China
JUNYAO LI, Guangdong University of Technology, China
KANGBO ZHAO, Guangdong University of Technology, China
SEN ZHANG, TikTok, ByteDance, Australia
YUKAI SHI*, Guangdong University of Technology, China
LIANG LIN, Sun Yat-sen University, China

Face sketch synthesis is a technique aimed at converting face photos into sketches. Existing face sketch synthesis research mainly relies on training with numerous photo-sketch sample pairs from existing datasets. However, these large-scale discriminative learning methods will have to face problems such as data scarcity and high human labor costs. Once the training data becomes scarce, their generative performance significantly degrades. In this paper, we propose a one-shot face sketch synthesis method based on diffusion models. We optimize text instructions on a diffusion model using face photo-sketch image pairs. Then, the instructions derived through gradient-based optimization are used for inference. To simulate real-world scenarios more accurately and evaluate method effectiveness more comprehensively, we introduce a new benchmark named One-shot Face Sketch Dataset (OS-Sketch). The benchmark consists of 400 pairs of face photo-sketch images, including sketches with different styles and photos with different backgrounds, ages, sexes, expressions, illumination, etc. For a solid out-of-distribution evaluation, we select only one pair of images for training at each time, with the rest used for inference. Extensive experiments demonstrate that the proposed method can convert various photos into realistic and highly consistent sketches in a one-shot context. Compared to other methods, our approach offers greater convenience and broader applicability. The dataset will be available at: https://github.com/HanWu3125/OS-Sketch

CCS Concepts: • **Human-centered computing** → **Social media**.

Additional Key Words and Phrases: Human Facial Sketch, One-Shot, Diffusion Model, Out-of-Distribution

ACM Reference Format:
Han Wu, Junyao Li, Kangbo Zhao, Sen Zhang, Yukai Shi, and Liang Lin. 2018. One-shot Face Sketch Synthesis in the Wild via Generative Diffusion Prior and Instruction Tuning. *J. ACM* 37, 4, Article 111 (August 2018), 17 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding author

Authors' Contact Information: Han Wu, 2112303125@mail2.gdut.edu.cn, Guangdong University of Technology, Guangzhou, Guangdong, China; Junyao Li, 2112403037@mail2.gdut.edu.cn, Guangdong University of Technology, Guangzhou, Guangdong, China; Kangbo Zhao, zhaokangbo329@gmail.com, Guangdong University of Technology, Guangzhou, Guangdong, China; Sen Zhang, senzhang.thu10@gmail.com, TikTok, ByteDance, Sydney, Australia; Yukai Shi, ykshi@gdut.edu.cn, Guangdong University of Technology, Guangzhou, Guangdong, China; Liang Lin, linliang@ieee.org, Sun Yat-sen University, Guangzhou, Guangdong, China.

ACM 1557-735X/2018/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

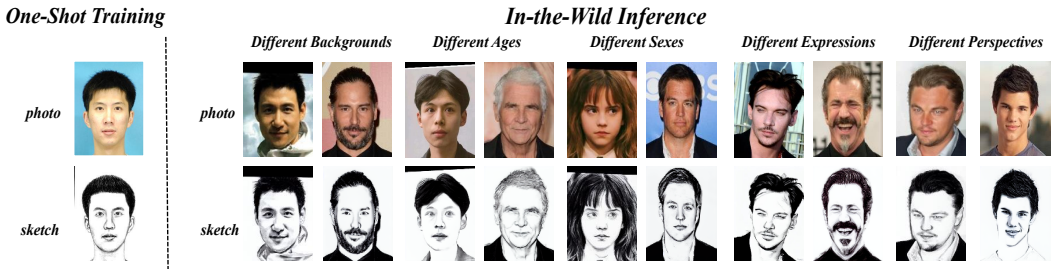

Fig. 1. An illustration of one-shot sketch synthesis. We call for a solid face sketch model is able to generate realistic sketches for face photos with different backgrounds, ages, sexes, expressions, perspectives, etc. Thus, in our model, only one pair of sketch samples is used for training, while a large number of diverse images are utilized for inference.

1 Introduction

Face sketch synthesis is a technique aimed at converting face photos into sketches, which holds significant potential for application in digital multimedia and security field [1, 4, 52]. As shown in Fig. 1, a variety of colorful photos are converted into realistic sketches. Existing face sketch synthesis researches [8, 10, 29, 31, 32] already have remarkable achievements. However, they mainly rely on training with numerous photo-sketch sample pairs from existing datasets. And these large-scale discriminative learning methods will have to face problems: data scarcity and high human labor costs.

As shown in Fig. 2, existing methods exhibit severe performance degradation when trained on just one photo-sketch pair. They even fail completely when confronted with out-of-distribution(OOD) photos. While traditional methods can achieve good generative capability as long as sufficient training data is available, the generalization ability often be violated when the training data is insufficient. In reality, the number of artists is limited. And the manual drawing of face sketches requires a substantial amount of time and effort. Consequently, hand-drawn sketches are costly and scarce. The limitation also results in the current situation where the number of existing datasets is limited. Nowadays, even the largest manually drawn face sketch dataset [8] contains only thousands of face sketch image pairs. How to balance the high manual sketching costs with the contradiction that traditional methods rely on massive data for training remains an unresolved issue.

In recent years, diffusion models [7, 18, 28, 34, 35] have become well-established tools for generating images. Especially for text-to-image diffusion models, which are capable of producing remarkable images when provided with an image and an editing instruction. By utilizing a text-to-image diffusion model for face sketch synthesis, it is possible to generate sketches without any training, which results in significant savings in training costs. Nevertheless, it is hard to achieve because relying on text descriptions solely is fragile. Specific textures and styles of sketches and the dispersed colour representation of different areas are difficult to describe through words. Recently, image editing via visual prompting has begun to be explored [2, 27, 43]. Visual Instruction Inversion (VISII) [27] introduces a robust framework for image editing based on visual prompting. By using images as prompts instead of solely text descriptions, image editing can be more intuitive and precise. Nonetheless, it mainly addresses basic style transfer applications for a general audience and does not explore the specific synthesis of face sketches.

In this paper, we introduce a new framework for face sketch synthesis called One-shot Face Sketch Synthesis. Inspired by image editing via visual prompting [27], we optimize text instruction on a diffusion model using only a single face photo-sketch image pairs. Then, the instruction

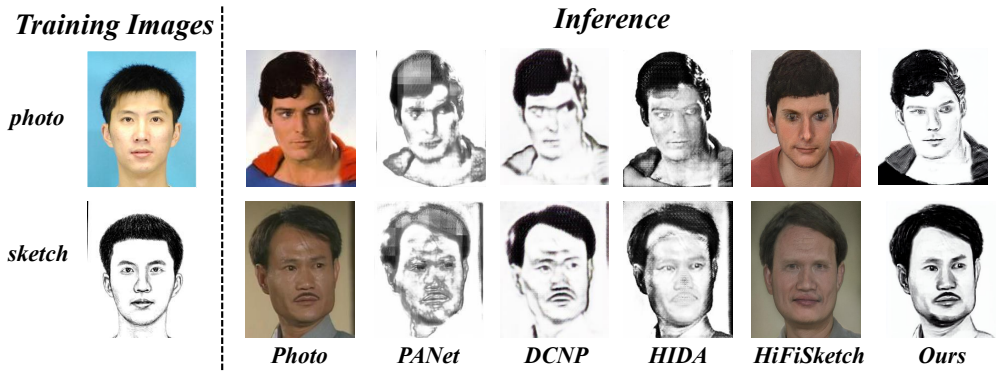

Fig. 2. A demonstration of Out-of-Distribution (OOD) challenge. When trained with only a single photo-sketch pair, existing methods exhibit severe performance degradation and even fail completely when confronted with out-of-distribution(OOD) photos. In contrast, our approach can convert various out-of-distribution (OOD) photos into sketches and maintain extremely high consistency with the photos.

derived through gradient-based optimization is used for inference. We verify the effectiveness of our method by processing various photos in a one-shot context. As shown in Fig. 1, we can generate realistic sketches with only a pair of images for training. Our method can handle photos in-the-wild with different backgrounds, ages, sexes, expressions, and perspectives. Our method can inherit the style of the training sketches while maintaining high consistency with the photos. At the same time, it significantly reduces energy costs. Compared to other methods, our approach offers greater convenience and broader applicability.

Moreover, we find that the existing dataset is limited by the variety of sketch styles. For example, existing datasets CUHK Face Sketch (CUFS) [42] dataset, CUHK Face Sketch FERET (CUFSF) [54] dataset and WildSketch [29] dataset only have a single sketch style. FS2K [8] dataset is the largest available hand-drawn face photo-sketch dataset currently, but it only has 3 sketch styles. However, in real-world scenarios, sketch styles have a greater variety of diversity. There are many exquisite sketches available on the internet. Besides, we believe using a dataset combining photographic features from diverse datasets would enable more comprehensive evaluation in one-shot face sketch synthesis research. In this paper, to more accurately simulate real-world scenarios and validate the effectiveness of our method, we introduce a new benchmark named One-shot Face Sketch Dataset (OS-Sketch). First we selected 40 sketches drawn by amateurs or bloggers along with corresponding photos from the internet. Then we select 100 photo-sketch pairs from the simple color backgrounds face sketch dataset [42], 110 pairs from the dark lighting backgrounds dataset [54] and 150 pairs from the in-the-wild dataset [29]. Our dataset possesses the following characteristics. First, our dataset comprises a diverse style of sketches. It not only includes sketches previously used in research but also incorporates out-of-distribution (OOD) style sketches selected from the internet. Our dataset is more complex than ever before and closer to real-world scenarios. Additionally, sketches selected from the internet are also available for researchers to conduct experimental studies. Second, the dataset not only contains traditional research photos with a single background but also includes numerous in-the-wild photos. It includes photos with different backgrounds, ages, sexes, expressions, perspectives and illumination. By exposing the model to a variety of photos in a one-shot context, its performance can be intuitively demonstrated.

To this end, the contributions of this paper can be summarized as follows:

- We present a new framework for face sketch synthesis. Specifically, only a pair of samples is required to ensure the achievement of face sketch synthesis in-the-wild under real-world conditions.
- We introduce a more challenging benchmark One-shot Face Sketch Dataset (OS-Sketch), which contains a total of 400 face photo-sketch image pairs, encompassing sketches with various styles and photos with different backgrounds, sexes, expressions, ages, illumination, perspectives, etc.
- Our work will significantly facilitate the development of face sketch tasks in real-world scenarios as former works focus on datasets with limited variation. Extensive experiments demonstrate that the proposed method can convert various photos into realistic and highly consistent sketches in a one-shot context.

2 RELATED WORK

2.1 Face Photo-Sketch Synthesis

Face photo-sketch synthesis is an important task in the field of computer vision [22, 41, 42], which aims to convert face photos into sketches. Traditional methods are mainly data-driven [41], relying on searching for similar blocks and linear combination weight computation. Sample-based approaches [21, 52, 58] have proven to be quite effective. And regression-based approaches [50, 57] produce synthesized images with more details. Nevertheless, traditional methods tend to be computationally intensive since they need to search through the entire training set to find the nearest neighbours.

In recent years, with the development of deep learning, many researchers have applied deep learning to face photo-sketch synthesis task [24, 56, 60]. For example, Zhang et al. [51] use an end-to-end fully convolutional network for the task of generating sketches from photos. Zhang et al. [49] further proposed a content-adaptive sketch generation method, which is achieved by using a branched fully convolutional network (BFCN) to learn multi-level and multi-scale decomposition representations of facial images. Jiao et al. [16] proposed a lightweight four-layer CNN method for generating face sketches, which can retain most of the details. Chen et al. [5] proposed a pyramid column feature based on CNN, which enriches the sketch textures and shadings. With the advancement of generative adversarial networks (GANs) [11, 37] and their remarkable success in diverse image generation tasks [6, 20, 46], researchers have begun applying GANs to face photo-sketch synthesis. For example, Philip et al. [32] attempted to apply conditional generative adversarial networks (cGANs) [15] to the synthesis of face sketches. Wang et al. [40] further refined their approach by combining the original cGANs [15] with a post-processing method called back-projection to produce more detailed sketches. Yi et al. [47] propose a composite GAN architecture called APDrawingGAN++, which consists of local networks to learn effective representations for specific facial features and a global network to capture the overall content. Yu et al. [48] introduced a new framework named Composition-Aided Generative Adversarial Network (CA-GAN), which leverages facial composition information to enhance the synthesis process, ensuring the authenticity of structure and consistency of texture in the generated facial sketches. Peng et al. [31] proposed a new cross-domain face photo-sketch synthesis framework named HiFiSketch. It learns to adjust the generator weights to achieve high-fidelity synthesis and manipulation, allowing for the translation of images between the photo domain and the sketch domain while enabling modifications according to the text input.

Despite achieving certain results, previous methods require extensive datasets for training, and obtain weak generalization capabilities on out-of-distribution cases. In this paper, we present a

one-shot face sketch synthesis method with generative diffusion prior, which fully addresses the generalization capabilities toward out-of-distribution cases.

2.2 Diffusion Models for Image Editing

Image editing is a technique that enables users to manipulate images based on their own expressions [19, 30]. Nowadays, diffusion models have garnered significant accomplishments in image generation [7, 34, 36, 38, 55, 61] and have been extended to image editing [3, 9, 12, 14, 26]. SDEdit [25] applies a pre-trained model to add noise to the input image and denoise it with a new user editing guide. Later, GLIDE [28] and Stable Diffusion models [34] incorporate this work into text-to-image generation, enabling effective text-based image editing. DiffusionCLIP [18] combines diffusion models with Contrastive Language-Image Pre-training (CLIP) [33] to perform zero-shot image manipulation and their method of image manipulation successfully performs both in the trained and unseen domain. Imagic [17] tried to leverage pre-trained diffusion models for image editing, but it necessitated fine-tuning of the model. In contrast, prompt-to-prompt [12] eliminates the need for training or fine-tuning of the model. It constitutes an intuitive image editing interface through editing only the textual prompt. Pix2pix-zero [30] introduces a diffusion-based image-to-image translation method that eliminates the need for training or prompts. Users simply specify the editing direction from the source to the target domain (such as cat → dog) without the need for manually crafting text prompts for the input image. Visual Instruction Inversion (VISII) [27] introduces a visual prompt-based approach to image editing, allowing users to replace hard-to-describe editing tasks with corresponding images.

Due to the powerful generative capabilities of diffusion models, in this paper we adopt a strategy that combines diffusion models with face sketch synthesis. Compared to previous methods, our method greatly reduces training costs while ensuring generation quality.

3 Methodology

In this section, we mainly introduce our method. First, we introduce a brief background on text-to-image diffusion models. Then, we present the training framework for One-shot Face Sketch Synthesis. The framework of our method is shown in Fig. 3.

3.1 Preliminaries

Diffusion models learn to generate data samples through a denoising sequence that estimates the score of the data distribution. In the forward process, Gaussian noise ϵ is gradually added to the input image x in T steps, producing a series of noisy samples x_1, ..., x_T. In the backward process, x_T will be restored to x through a denoising neural network.

Initially, diffusion models were operated in pixel space [7]. To enable diffusion models training on limited computational resources while retaining their quality and flexibility, Latent Diffusion Model (LDM) [34] proposes running the diffusion process in latent space. Latent Diffusion Model (LDM) [34] introduces an autoencoder model for image reconstruction. The encoder $\mathcal{E}$ encodes the input image into a 2D latent space z to obtain the latent image z_x. And the decoder $\mathcal{D}$ is used for decoding. To enable the diffusion model to generate more accurately, LDM [34] also introduces a domain-specific encoder τ_θ, which projects text prompts into an intermediate representation C_{Text}. Then, C_{Text} is inserted into the layers of the denoising network through cross-attention as an instruction of the model. Thus, the diffusion model is allowed to take text prompts as conditional input. The objective function is:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(x), C_{Text}, \epsilon \sim \mathcal{N}(0,1), t} \left\| \epsilon - \epsilon_\theta \left(z_{x_t}, t, C_{Text} \right) \right\|_2 \tag{1}$$

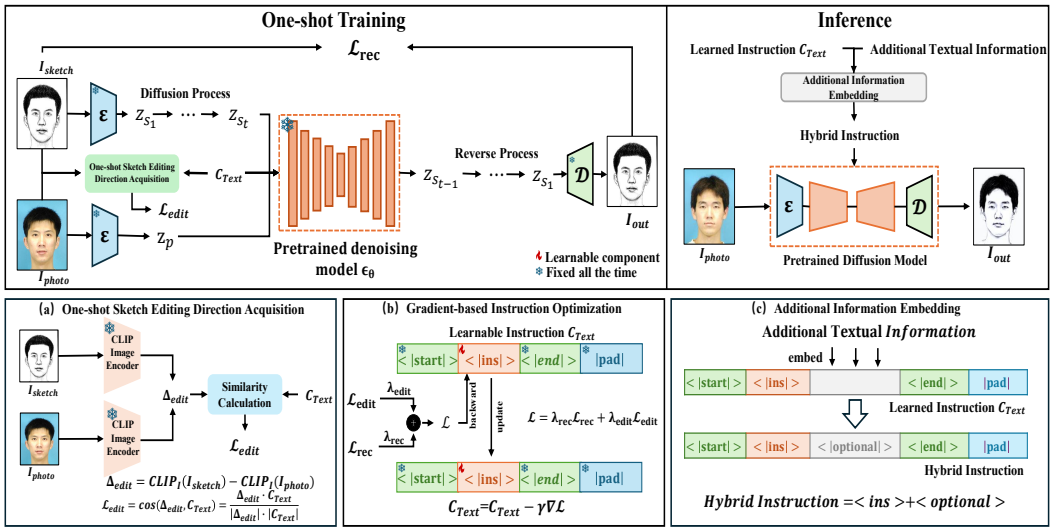

Fig. 3. Our framework jointly optimizes the text instruction C_{Text} through the image reconstruction strategy of a diffusion model for one-shot sketch synthesis. Here, $\mathcal{E}$ denotes the encoder, and $\mathcal{D}$ denotes the decoder. (a) We encode a pair of sketch samples $\{I_{photo}, I_{sketch}, \}$ into the embedding space using CLIP [33]. Then, we calculate the CLIP embedding distance $\triangle_{edit}$ between I_{photo} and I_{sketch}, and utilize $\triangle_{edit}$ to optimize the text instruction C_{Text}. In this way, C_{Text} is able to represent the editing direction from a photo to a sketch. We choose cosine similarity for the optimization process. (b) We perform gradient-based optimization on the learnable Instruction C_{Text}. We only optimize a part of the instruction, marked as $< ins >$. We use two losses $\mathcal{L}_{rec}$ and $\mathcal{L}_{edit}$ to jointly optimize C_{Text}. (c) During inference, we can embed additional textual information into learned instruction C_{Text}. The embedded part is marked as $< optional >$. The final hybrid instruction is formulated by integrating the optimized $< ins >$ with the supplementary $< optional >$ provided more historical context.

where t denotes each time step.

However, text instructions C_{Text} are hard to align with the desired edits precisely. This limitation causes diffusion-generated images to struggle to fully adhere to the pixel-level details of the input images. To address this challenge, InstructPix2Pix [3] proposes incorporating the input image into the denoising network. Specifically, the input image x is encoded into a conditional image $\mathcal{E}(x)$ and concatenated with the latent image z_{x_t}. As a result, the objective function Eq. 1 is modified to:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(x), C_{Text}, \mathcal{E}(x), \epsilon \sim \mathcal{N}(0,1), t} \left\| \epsilon - \epsilon_\theta \left(z_{x_t}, t, C_{Text}, \mathcal{E}(x) \right) \right\|_2 \tag{2}$$

3.2 Framework

Existing methods require training on large datasets to generate face sketches, which leads to higher training costs. However, there are only a limited number of artists and face sketches available in most real-world scenarios. To address the issue of high training costs, we propose a one-shot face sketch synthesis method based on diffusion models. As shown in Fig. 4, to model real-world challenges, we only use one pair of face photo-sketch samples for training, and then apply the other samples for inference. *Our method is capable of handling a diverse range of photos, including photos in-the-wild.*

The primary objective of our method is to investigate a text instruction C_{Text} that can accurately describe the editing direction from a face photo to a specific sketch. By using the sketch instruction,

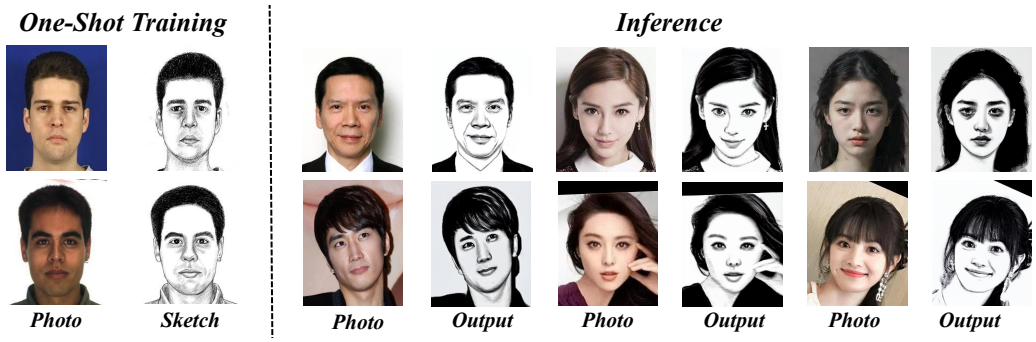

Fig. 4. An illustration of the proposed in-the-wild face sketch algorithm. In our framework, only one pair of face photo-sketch samples is required for training, and the rest images will be fine for testing. Confronting with photos featuring different expressions, backgrounds, and sex, our method is capable of handling a diverse range of photos in-the-wild.

our model is enable to generate the expected results with in-the-wild input. Specifically, we set a pair of samples as $\{I_{photo}, I_{sketch}\}$, where I_{photo} is the face photo and I_{sketch} is the sketch. We perform gradient-based optimization on the instruction C_{Text} by leveraging these specific images with rich visual information. To this end, our method is mainly divided into two stages: (1) Optimization through image reconstruction strategies. (2) Optimization through the image editing direction determined.

Face Reconstruction with Diffusion Model. Usually, diffusion models are trained on a sequence of gradually noisier images over a series of timesteps, e.g., t = 1, . . . , T. With the supervision of the objective function, the diffusion models can acquire the powerful ability to reconstruct images from noise. In this paper, we adopt the image reconstruction strategies of diffusion models to optimize the instruction C_{Text}. In this way, C_{Text} is able to encompass the description to recover images from noise, which is a basic and important function of diffusion models for generating images.

Specifically, we conduct training based on a denoising network. Inspired by Eq. 2, given a pair of sketch samples $\{I_{photo}, I_{sketch}\}$, we encode and add noise to the sketch I_{sketch}. At the same time, we encode the face photo I_{photo} and add it to the denoising network to ensure that the model can follow the pixel-level information of the input photo I_{photo}. Then, without fine-tuning the model, we train the noise prediction ability of the model by training the text instruction C_{Text}.

Hence, for the sketch sample pair $\{I_{photo}, I_{sketch}\}$, we define the image reconstruction loss $\mathcal{L}_{rec}$ as follows:

$$\mathcal{L}_{rec} = \mathbb{E}_{\mathcal{E}(I_{sketch}),c_{Text},z_p,\epsilon\sim\mathcal{N}(0,1),t} \left\| \epsilon - \epsilon_\theta \left(z_{s_t}, t, C_{Text}, z_p \right) \right\|_2 \tag{3}$$

where z_{s_t} represents the denoised variable of the sketch I_{sketch} after encoding and adding noise at timestep t, and z_p represents the latent variable after encoding the photo I_{photo} and adding it to the denoising network.

One-shot Sketch Editing Direction Acquisition. When given an editing instruction, the model can edit the input image according to the editing direction provided by the instruction. However, if C_{Text} is only optimized based on Eq. 3, it can only learn the prior information inherent to the reconstructed image. It cannot accurately describe the editing direction from a photo to a sketch, which may affect the generation performance of the model.

Algorithm 1 One-shot Face Sketch Instruction Acquisition, given a pre-trained denoising model ϵ_θ, an encoder $\mathcal{E}$ and a CLIP image encoder $CLIP_I(\cdot)$.

Input: A sketch sample pair $\{I_{photo}, I_{sketch}\}$, encoded as z_p, z_s; Optimization steps N; Timesteps T; Initialized text instruction C_{Text}; CLIP embedding distance(between I_{photo} and I_{sketch}) $\triangle_{edit}$; Hyperparameters λ_{rec} and λ_{edit}; Learning rate γ
Output: Text instruction C_{Text}
Optimize text instruction C_{Text}
for $i \leftarrow 1$ to N **do**
  Sample t from $U(0, T)$; ϵ from $N(0,1)$
  $z_{s_t} \leftarrow z_s+\epsilon$ at timestep t
  $\mathcal{L}=\lambda_{rec} \left\|\epsilon - \epsilon_\theta \left(z_{s_t}, t, c_{Text}, z_p\right)\right\|_2+\lambda_{edit} cos(\triangle_{edit}, C_{Text})$
  $C_{Text} \leftarrow C_{Text}$ - $\gamma \nabla \mathcal{L}$
end for
return C_{Text}

Nowadays, Contrastive Language-Image Pre-training (CLIP) [33] has become a robust tool for image editing as it can establish a connection between texts and images. As shown in Fig. 3(a), we calculate the distance between photos and sketches in the high-dimensional space. We use the distance to optimize C_{Text}, enabling it to learn the distinction representation between photos and sketches. Thus, C_{Text} becomes to possess a representation of the editing direction from photo to sketch.

Specifically, given a sample pair $\{I_{photo}, I_{sketch}\}$, we calculate the distance between the CLIP embeddings of I_{photo} and I_{sketch}, denoted as $\triangle_{edit}$. Therefore, $\triangle_{edit}$ can be expressed as:

$$\triangle_{edit} = CLIP_I(I_{sketch}) - CLIP_I(I_{photo}) \tag{4}$$

where $CLIP_I(\cdot)$ represents the image encoder of CLIP.

Then we use $\triangle_{edit}$ to optimize the text instruction C_{Text}. We choose cosine similarity as the loss function of image editing, denoted as $\mathcal{L}_{edit}$. Therefore, $\mathcal{L}_{edit}$ can be expressed as:

$$\mathcal{L}_{edit} = cos(\triangle_{edit}, C_{Text}) = \frac{\triangle_{edit} \cdot C_{Text}}{|\triangle_{edit}|| C_{Text}|} \tag{5}$$

where cos(·) denotes the cosine function.

Finally, we combine the image reconstruction loss $\mathcal{L}_{rec}$ (Eq. 3) and the image editing loss $\mathcal{L}_{edit}$ (Eq. 4) to jointly optimize the instruction C_{Text}:

$$\mathcal{L} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{edit}\mathcal{L}_{edit} \tag{6}$$

where λ_{rec} and λ_{edit} are hyperparameters.

In this way, after gradient-based optimization, C_{Text} can not only follow the pixel-level details of the input image but also summarize the information of the sketch image while containing the editing direction from the photo to the sketch. The complete algorithm for acquiring One-shot Face Sketch Instruction C_{Text} is shown in Algorithm 1.

Hybrid Instruction Tuning. Once the gradient-based optimized C_{Text} is obtained, we can use it as the input textual instruction for the diffusion model during inference. To control the generation direction of the model more precisely, our method allows for additional information to be embedded during inference to guide the process. As shown in Fig. 3(b), during training, we only optimize a portion of C_{Text}. The optimized part is marked as $< ins >$. Then, during inference, we can manually input additional text information to embed into C_{Text}, as shown in Fig. 3(c). The

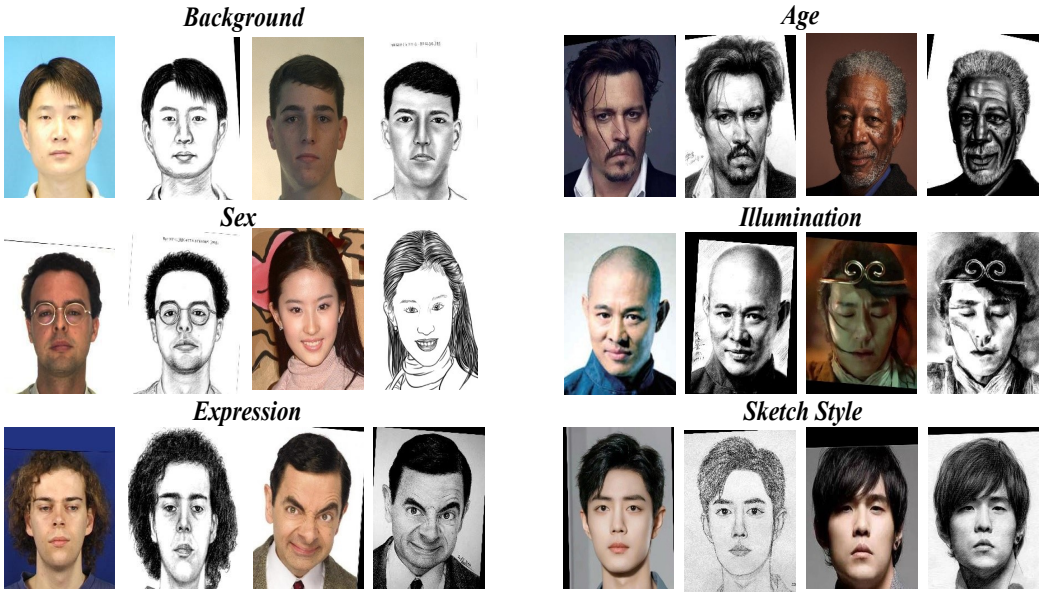

Fig. 5. An illustration of OS-Sketch dataset. Our dataset contains photos with various backgrounds, sexes, expressions, ages, illumination as well as sketches in various styles.

embedded part is marked as $< optional >$. The final hybrid instruction used for inference can be expressed as:

$$Hybrid \quad Instruction =< ins > + < optional > \tag{7}$$

where $< ins >$ is the optimized part of C_{Text}, and $< optional >$ is the additional part embedded to provide more historical context during inference.

Eventually, we can use the hybrid instruction composed of the learned C_{Text} and the embedded additional information as the text condition for the pre-trained diffusion model. With the allowance of embedding additional information into the instruction, we can control the direction of image generation more flexibly.

4 Experiment

In this section, we first introduce the implementation details of our method. Then, we describe the datasets and evaluation metrics. Next, we present the experimental results from both qualitative and quantitative perspectives, verifying the effectiveness of the proposed method. Finally, we provide ablation experiments to validate the effectiveness of each module.

4.1 Implementation Details and Datasets

In this paper, we use [45] to automatically generate a text description from the input sketch, serving as the initialized C_{Text}. We use a frozen pre-trained model to optimize the instruction C_{Text} at T=1000 timesteps, with the optimization step N set to 12000. We use the AdamW optimizer [23] with learning rate γ=0.001, λ_{rec}=4 and λ_{edit}=0.1. All experiments were conducted on a machine with an NVIDIA 3090 24G GPU.

One-shot Face Sketch Dataset. To more accurately simulate real-world scenarios and examine the feasibility of sketch synthesis methods in a one-shot context, we introduce a new benchmark

Table 1. Quantitative results on the OS-Sketch. The results of the top two performance are highlighted in red and blue, respectively.

Method	*SSIM* ↑	*LPIPS* ↓	*FID* ↓
PANet [29]	0.427	0.446	116.67
DCNP [59]	**0.448**	0.495	111.10
HIDA [10]	0.408	**0.412**	79.13
HiFiSketch [31]	0.354	0.497	128.62
InstructPix2Pix [3]	0.437	**0.348**	**78.91**
PnP-Diffusion [39]	0.388	0.545	139.59
pix2pix-zero [30]	0.336	0.632	190.25
Ours	**0.461**	**0.348**	**68.46**

called One-shot face sketch datasets (OS-Sketch). As shown in Fig. 5, our dataset contains 400 face photo-sketch image pairs, encompassing sketches with various styles and photos with different backgrounds, sexes, expressions, ages, illumination, etc. Our dataset features the most complex range of sketch styles and is the first to be used for one-shot face sketch synthesis, which comprises 100 face photo-sketch image pairs from the simple color backgrounds face sketch dataset [42], 110 pairs from the dark lighting backgrounds dataset [54], 150 pairs from the in-the-wild dataset [29] and 40 pairs selected from the internet. The dataset not only contains traditional research photos with a single background but also includes various in-the-wild photos.*We select only ONE pair of images for training, with the REST used for inference.* We assess the effectiveness of the sketch synthesis method by challenging it to process a diverse range of photos and sketches in a one-shot context. During the inference, the model will encounter numerous independent and identically distributed (i.i.d) and out-of-distribution (OOD) photos. Compared to previous datasets, we believe ours enables more comprehensive evaluation of method performance. All photos and sketches are highly aligned. And all sketches are uniformly converted into high-contrast grayscale images during evaluation.

Evaluation Metrics: We have selected three evaluation metrics to verify the effectiveness of our method, namely Structural Similarity Index (SSIM)[44], Learned Perceptual Image Patch Similarity (LPIPS)[53], and Fréchet Inception Distance (FID)[13].

4.2 Comparison

To demonstrate the superiority of our method, we compared it with advanced methods that perform well in face sketch synthesis such as PANet[29], DCNP[59], HIDA[10] and HiFiSketch[31]. Furthermore, to highlight the advantages of the one-shot strategy, we also conduct qualitative and quantitative comparisons with the text-driven diffusion-based style transfer techniques InstructPix2Pix [3], PnP-Diffusion[39] and pix2pix-zero [30] on the OS-Sketch. For a fair comparison, we only used one pair of face photo-sketch samples for training for all methods. In order to control the generation direction of the model more precisely, we select "covert the image color to black and white with a white background" as additional embedded textual information during inference.

Quantitative and Qualitative Comparisons. The quantitative results on the OS-Sketch are shown in Table. 1. We achieve the best results. Our method significantly improves upon previous methods in terms of Structural Similarity Index (SSIM)[44], Learned Perceptual Image Patch Similarity[53], and Fréchet Inception Distance (FID)[13] metrics. Traditional methods require large datasets for training. In a one-shot context, the models are limited by their training capacity. They can only learn to convert one photo into a sketch with one background, one face, one tone and

Fig. 6. Qualitative comparisons. We conduct a qualitative comparison of our method with other methods on OS-Sketch. Compared to existing methods, our method has better results.

one style of sketch. When switching to other diverse face photos, there is a substantial decline in the predictive capability of the models. They even fail completely when confronted with out-of-distribution(OOD) photos. As for the text-driven diffusion-based methods, they rely on text descriptions solely. However, text descriptions sometimes can be ambiguous, making it difficult to generate precise and appropriate results. Therefore, the results of typical face sketch methods are not satisfactory either. Our method maintains good generation results with one-shot training, *which significantly speeds up the training efficiency, reduces energy costs, and addresses the issue of dataset scarcity.*

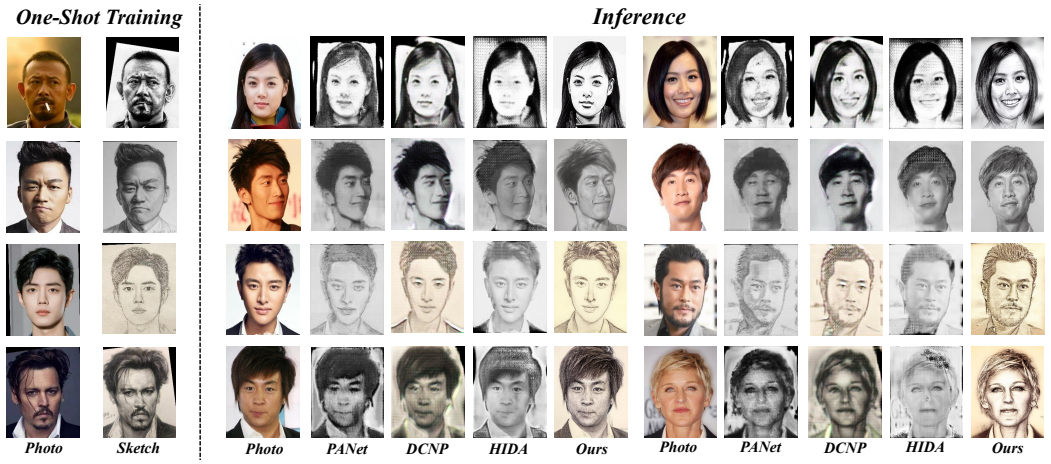

Fig. 7. We exhibit an in-the-wild experiment by using 4 challenging samples selected from the internet for out-of-distribution sketch synthesis. Experiments demonstrate that our method achieves good performance for sketches in-the-wild. It not only indicates that we can directly conduct research on sketches in-the-wild instead of relying on existing datasets, but also means that we can manipulate the style of sketches more flexibly.

The qualitative results of OS-Sketch are shown in Fig. 6. In a one-shot context, most existing methods fail. The sketches generated by them have issues with structural distortion, colour displacement and blurriness. For the text-driven diffusion-based methods, the generation results are also unsatisfactory due to the difficulty in describing specific sketch generation tasks with solely text descriptions. The generated sketches have issues with facial structure, expressions, and regional colours, which are difficult to control precisely through words. Compared to existing methods, our method has better results. In a one-shot scene, our method can generate complete and realistic sketches from face photos with different backgrounds, sexes, illumination, etc. Our approach can convert both independent and identically distributed (i.i.d) and out-of-distribution (OOD) photos into sketches at the same time. It remains effective even when applied to photos with complex backgrounds. In contrast, traditional methods sometimes even fail to generate a complete face. The experiments demonstrate the superiority of our method, which achieves impressive results with only one pair of samples for training, significantly improving experimental efficiency.

Efficiencies of Training and Inference Phrases. Moreover, we conduct an efficiency analysis of our method compared to other approaches. Specifically, we compared the training efficiency of our method with the state-of-the-art face sketch synthesis techniques and compared the inference efficiency with the diffusion-based methods. Both the training and inference comparisons are conducted using a single image. As shown in Table. 2, compared to PANet [29], DCNP [59], HIDA [10] and HiFiSketch [31], our method not only achieves top-tier training speed but also delivers superior generation results in terms of LPIPS [53] and FID [13]. As shown in Table. 3, compared to PnP-Diffusion [39] and pix2pix-zero [30], although they do not require any training, our inference speed surpasses theirs. And our method generates results with greater precision. The experimental results confirm that our method offers higher accuracy and practical usability.

Out-of-Distribution Comparison. OS-Sketch contains 40 photos selected from the internet and corresponding sketches drawn by bloggers or amateurs. To further explore our method, we

Table 2. Comparison of *training* efficiency. The best results are shown in bold and suboptimal results are underlined.

Method	PANet [29]	DCNP [59]	HIDA [10]	HiFiSketch [31]	Ours
Training times (S)	**6062.20**	11954.76	10361.46	71058.99	<u>7616.35</u>
LPIPS ↓	0.446	0.495	<u>0.412</u>	0.497	**0.348**
FID ↓	116.67	111.10	<u>79.13</u>	128.62	**68.46**

Table 3. Comparison of *inference* efficiency. The best results are shown in bold.

Method	PnP-Diffusion [39]	pix2pix-zero [30]	Ours
Inference times (S)	194.21	24.93	**5.99**
LPIPS ↓	0.545	0.632	**0.348**
FID ↓	139.59	190.25	**68.46**

select 4 original sketches with distinctive characteristics for training and testing the rest images. The qualitative results are shown in Fig. 7.

As can be seen from the visualization results, our method is capable of producing sketches with reasonable quality even with face photo-sketch pairs selected from the internet. Our method can generate realistic sketch textures and maintains a tone similar to the training samples, whereas other methods fail in a one-shot context. Experiments demonstrate the superiority of our method. In a one-shot context, our method can still adapt to out-of-distribution (OOD) styles of sketches. This attempt also suggests that in future research, specialized datasets are not required any more. Instead, we can directly select random face photo-sketch pairs from the internet. This finding greatly reduces the dependence on datasets and allows us to control the style of the sketches more flexibly.

4.3 Ablation Study

Effects of Generative Diffusion Prior. As shown in Fig. 8, we demonstrate the difference in effect between with and without $\mathcal{L}_{edit}$. Without the use of $\mathcal{L}_{edit}$, the learned C_{Text} lacks information on the editing direction from a photo to a sketch. The model only learns the information of the sketch image itself and does not understand the expected editing direction. Therefore, the generated image has not been truly converted into a sketch. Thus, it is essential to incorporate $\mathcal{L}_{edit}$ to obtain a more satisfactory editing direction.

Effects of Hybrid Instruction Tuning. As shown in Fig. 9, we demonstrate the difference in effect between guidance with hybrid instruction and with solely $< ins >$ without additional information input. It is clear that without manually inputting extra information, the tone of the images generated by the model sometimes may have some deviations. When selecting "convert the image color to black and white, white background" as the additional embedded textual information (i.e., <optional>), the model generates the desired results. The experiments confirm that with the guidance of hybrid instruction tuning, the model can better control the generation effect and accurately convert a face photo into the desired sketch.

5 Limitation and Conclusion

In this paper, we introduce a new framework for face sketch synthesis called One-shot Face Sketch Synthesis. To more accurately simulate real-world scenarios and evaluate the feasibility of our method in a one-shot context, we introduce a new benchmark called OS-Sketch. It includes sketches of various styles and photos with different backgrounds, expressions, sexes, illumination,

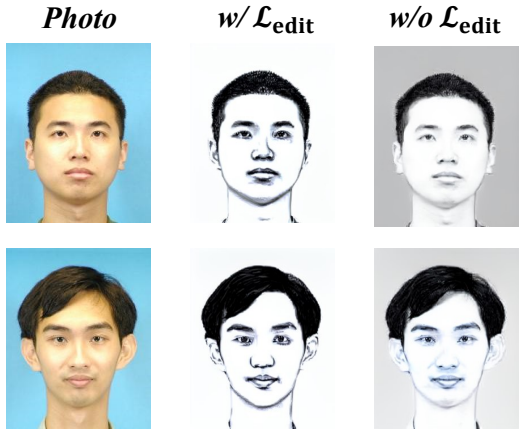

Fig. 8. Comparison of the generation results of "w/ $\mathcal{L}_{edit}$" and "w/o $\mathcal{L}_{edit}$". Without the use of $\mathcal{L}_{edit}$, the generated image has not been truly converted into a sketch image. Experiments demonstrate the effects of sketch prompt acquisition.

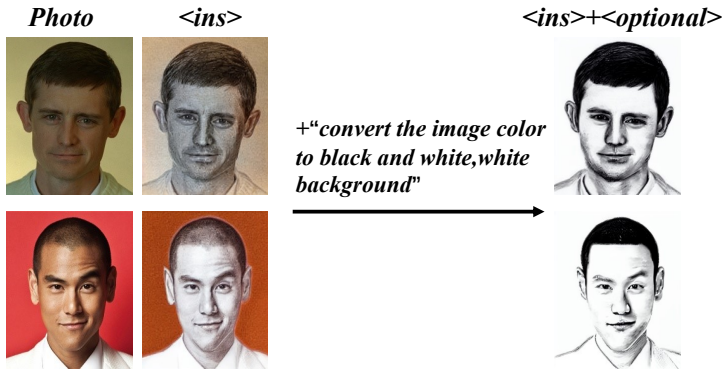

Fig. 9. Comparison of the generation results of guidance with hybrid instruction and with solely < *ins* > . Experiments demonstrate that instruction with embedded additional information have a more accurate result.

etc. Extensive experiments demonstrate the excellent performance of our framework in both quantitative and qualitative results. Our method can generate realistic sketches and achieve face sketch synthesis in-the-wild using only one pair of samples for training. In one-shot scenes, our method maintains high consistency between the sketches and the original photos. Additionally, our method enables the model to generate sketches in corresponding styles by training on sketches selected from the internet. Our method is highly suitable for face sketch synthesis, significantly reducing training costs and improving training efficiency.

However, our method is based on the text-to-image diffusion model, which still requires corresponding prompt and instructions. Without additional textual information embeddings, sometimes we may not have the expected results. Besides, there may be limitations in handling all types of sketch styles or achieve more precise sketch textures due to the operational mode of text-to-image diffusion models. In the future, more advanced solutions should be proposed to further enhance efficiency, optimize textures or other details.

References

[1] Swapna Agarwal and Dipti Prasad Mukherjee. 2018. Synthesis of realistic facial expressions using expression map. *IEEE Transactions on Multimedia* 21, 4 (2018), 902–914.

[2] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. 2022. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems* 35 (2022), 25005–25017.

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

[4] Eva Cetinic and James She. 2022. Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 2 (2022), 1–22.

[5] Chaofeng Chen, Xiao Tan, and Kwan-Yee K Wong. 2018. Face sketch synthesis with style transfer using pyramid column feature. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 485–493.

[6] Zijun Deng, Xiangteng He, and Yuxin Peng. 2023. LFR-GAN: local feature refinement based generative adversarial network for text-to-image generation. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 6 (2023), 1–18.

[7] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

[8] Deng-Ping Fan, Ziling Huang, Peng Zheng, Hong Liu, Xuebin Qin, and Luc Van Gool. 2022. Facial-sketch synthesis: A new challenge. *Machine Intelligence Research* 19, 4 (2022), 257–287.

[9] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*. Springer, 89–106.

[10] Fei Gao, Yifan Zhu, Chang Jiang, and Nannan Wang. 2023. Human-Inspired Facial Sketch Synthesis with Dynamic Adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7237–7247.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control.(2022). *URL https://arxiv. org/abs/2208.01626* (2022).

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[14] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. 2023. ReVersion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495* (2023).

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

[16] Licheng Jiao, Sibo Zhang, Lingling Li, Fang Liu, and Wenping Ma. 2018. A modified convolutional neural network for face sketch synthesis. *Pattern Recognition* 76 (2018), 125–136.

[17] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.

[18] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2426–2435.

[19] Bo Li, Xiao Lin, Bin Liu, Zhi-Fen He, and Yu-Kun Lai. 2023. Lightweight text-driven image editing with disentangled content and attributes. *IEEE Transactions on Multimedia* (2023).

[20] Yongkang Li, Qifan Liang, Zhen Han, Wenjun Mai, and Zhongyuan Wang. 2024. Few-shot face sketch-to-photo synthesis via global-local asymmetric image-to-image translation. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 10 (2024), 1–24.

[21] Decheng Liu, Jie Li, Nannan Wang, Chunlei Peng, and Xinbo Gao. 2018. Composite components-based face sketch recognition. *Neurocomputing* 302 (2018), 46–54.

[22] Qingshan Liu, Xiaoou Tang, Hongliang Jin, Hanqing Lu, and Songde Ma. 2005. A nonlinear approach for face sketch synthesis and recognition. In *2005 IEEE Computer Society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. IEEE, 1005–1010.

[23] I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[24] Dan Lu, Zhenxue Chen, QM Jonathan Wu, and Xuetao Zhang. 2019. FCN based preprocessing for exemplar-based face sketch synthesis. *Neurocomputing* 365 (2019), 113–124.

[25] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021).

[26] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6038–6047.

[27] Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. 2024. Visual instruction inversion: Image editing via image prompting. *Advances in Neural Information Processing Systems* 36 (2024).

[28] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

[29] Lin Nie, Lingbo Liu, Zhengtao Wu, and Wenxiong Kang. 2022. Unconstrained face sketch synthesis via perception-adaptive network and a new benchmark. *Neurocomputing* 494 (2022), 192–202.

[30] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

[31] Chunlei Peng, Congyu Zhang, Decheng Liu, Nannan Wang, and Xinbo Gao. 2023. HiFiSketch: High Fidelity Face Photo-Sketch Synthesis and Manipulation. *IEEE Transactions on Image Processing* (2023).

[32] Chikontwe Philip and Lee Hyo Jong. 2017. Face sketch synthesis using conditional adversarial networks. In *2017 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 373–378.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[35] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*. 1–10.

[36] Bowen Sun, Guo Lu, and Shibao Zheng. 2023. DiFace: Cross-Modal Face Recognition through Controlled Diffusion. *ACM Transactions on Multimedia Computing, Communications and Applications* (2023).

[37] Hamidou Tembine. 2019. Deep learning meets game theory: Bregman-based algorithms for interactive deep generative adversarial networks. *IEEE transactions on cybernetics* 50, 3 (2019), 1132–1145.

[38] Jiahang Tu, Wei Ji, Hanbin Zhao, Chao Zhang, Roger Zimmermann, and Hui Qian. 2025. Driveditfit: Fine-tuning diffusion transformers for autonomous driving data generation. *ACM Transactions on Multimedia Computing, Communications and Applications* 21, 3 (2025), 1–29.

[39] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1921–1930.

[40] Nannan Wang, Wenjin Zha, Jie Li, and Xinbo Gao. 2018. Back projection: An effective postprocessing method for GAN-based face sketch synthesis. *Pattern Recognition Letters* 107 (2018), 59–65.

[41] Nannan Wang, Mingrui Zhu, Jie Li, Bin Song, and Zan Li. 2017. Data-driven vs. model-driven: Fast face sketch synthesis. *Neurocomputing* 257 (2017), 214–221.

[42] Xiaogang Wang and Xiaoou Tang. 2008. Face photo-sketch synthesis and recognition. *IEEE transactions on pattern analysis and machine intelligence* 31, 11 (2008), 1955–1967.

[43] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. 2023. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6830–6839.

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[45] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems* 36 (2024).

[46] Xian Wu, Kun Xu, and Peter Hall. 2017. A survey of image synthesis and editing with generative adversarial networks. *Tsinghua Science and Technology* 22, 6 (2017), 660–674.

[47] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. 2019. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10743–10752.

[48] Jun Yu, Xingxin Xu, Fei Gao, Shengjie Shi, Meng Wang, Dacheng Tao, and Qingming Huang. 2020. Toward realistic face photo–sketch synthesis via composition-aided GANs. *IEEE transactions on cybernetics* 51, 9 (2020), 4350–4362.

[49] Dongyu Zhang, Liang Lin, Tianshui Chen, Xian Wu, Wenwei Tan, and Ebroul Izquierdo. 2016. Content-adaptive sketch portrait generation by decompositional representation learning. *IEEE transactions on image processing* 26, 1 (2016),

328–339.

[50] Jiewei Zhang, Nannan Wang, Xinbo Gao, Dacheng Tao, and Xuelong Li. 2011. Face sketch-photo synthesis based on support vector regression. In *2011 18th IEEE International Conference on Image Processing*. IEEE, 1125–1128.

[51] Liliang Zhang, Liang Lin, Xian Wu, Shengyong Ding, and Lei Zhang. 2015. End-to-end photo-sketch generation via fully convolutional representation learning. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. 627–634.

[52] Mingjin Zhang, Nannan Wang, Yunsong Li, and Xinbo Gao. 2019. Deep latent low-rank representation for face sketch synthesis. *IEEE transactions on neural networks and learning systems* 30, 10 (2019), 3109–3123.

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

[54] Wei Zhang, Xiaogang Wang, and Xiaoou Tang. 2011. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR 2011*. IEEE, 513–520.

[55] Mohan Zhou, Yalong Bai, Qing Yang, and Tiejun Zhao. 2024. StyleInject: Parameter Efficient Tuning of Text-to-Image Diffusion Models. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).

[56] Mingrui Zhu, Jie Li, Nannan Wang, and Xinbo Gao. 2019. A deep collaborative framework for face photo–sketch synthesis. *IEEE transactions on neural networks and learning systems* 30, 10 (2019), 3096–3108.

[57] Mingrui Zhu and Nannan Wang. 2016. A simple and fast method for face sketch synthesis. In *Proceedings of the International Conference on Internet Multimedia Computing and Service*. 168–171.

[58] Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Li. 2017. Deep graphical feature learning for face sketch synthesis. In *Proceedings of the 26th international joint conference on artificial intelligence*. 3574–3580.

[59] Mingrui Zhu, Zicheng Wu, Nannan Wang, Heng Yang, and Xinbo Gao. 2023. Dual conditional normalization pyramid network for face photo-sketch synthesis. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 9 (2023), 5200–5211.

[60] Zhu, Mingrui and Li, Jie and Wang, Nannan and Gao, Xinbo. 2020. Knowledge distillation for face photo–sketch synthesis. *IEEE Transactions on Neural Networks and Learning Systems* 33, 2 (2020), 893–906.

[61] Kaifeng Zou, Sylvain Faisan, Boyang Yu, Sébastien Valette, and Hyewon Seo. 2024. 4d facial expression diffusion model. *ACM Transactions on Multimedia Computing, Communications and Applications* 21, 1 (2024), 1–23.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009