Controllable Sketch-to-Image Translation for Robust Face Synthesis

Shuai Yang, *Member, IEEE*, Zhangyang Wang, *Senior Member, IEEE*, Jiaying Liu, *Senior Member, IEEE*, and Zongming Guo, *Member, IEEE*

***Abstract*— In this paper, we propose a novel controllable sketch-to-image translation framework that allows users to interactively and robustly synthesize and edit face images with hand-drawn sketches. Inspired by the coarse-to-fine painting process of human artists, we propose a novel dilation-based sketch refinement method to refine sketches at varied coarse levels without the need for real sketch training data. We further investigate multi-level refinement that enables users to flexibly define how "reliable" the input sketch should be considered for the final output through a refinement level control parameter, which helps balance between the realism of the output and its structural consistency with the input sketch. It is realized by leveraging scale-aware style transfer to model and adjust the style features of sketches at different coarse levels. Moreover, advanced user controllability in terms of the editing region control, facial attribute editing, and spatially non-uniform refinement is further explored for fine-grained and semantic editing. We demonstrate the effectiveness of the proposed method in terms of visual quality and user controllability through extensive experiments including qualitative and quantitative comparison with state-of-the-art methods, ablation studies and various applications.**

***Index Terms*— Face synthesis, sketch-to-image translation, user control, image editing.**

I. Introduction

HAND-DRAWN sketches are highly succinct yet expressive representations for humans to depict real-world objects. At the same time, sketches are easy to draw and edit, especially with the popularity of touch screen devices. Therefore, sketching has become one of the important means for people to show their ideas. Accordingly, sketch-based image synthesis has been studied a lot for generating and editing photo-realistic images based on the structural guidance of sketches. It allows normal users to create novel images or modify photos by simply drawing several lines as shown in Fig. 1, without the need to carefully handle the complex photos themselves.

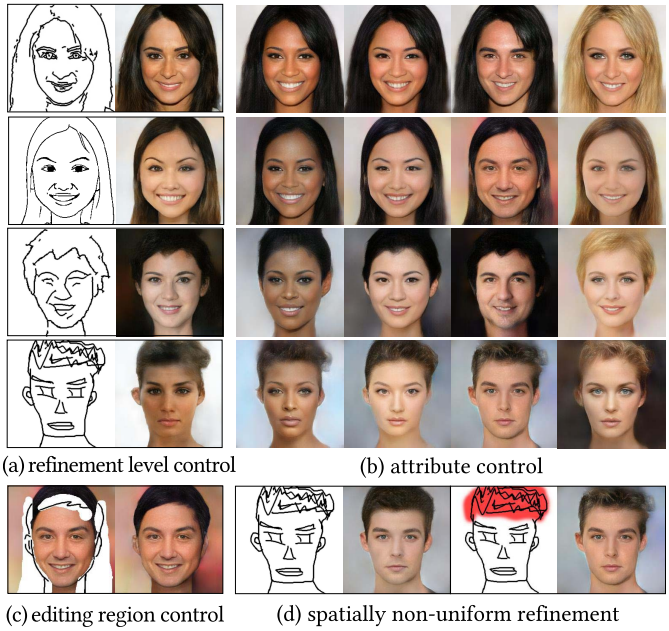


Fig. 1. Our model allows users to synthesize and edit facial images based on hand-drawn sketches. (a) Our model works robustly on (from top to bottom) edge map, fine sketch, rough sketch and poor sketch by setting refinement level ℓ adaptive to the quality of the input, *i.e.*, higher ℓ for poorer sketches. (b) Our model enables users to select facial attributes such as (from left to right) race, gender, and hair color. (c) Users can provide masks to specify the editing regions of an image. (d) Our model supports spatially non-uniform sketch refinement for more flexible controllability. In this case, the red region uses a lower refinement level to preserve the hair details.

Since pairs of hand-drawn sketches and photos are expensive and tedious to collect, previous methods [1]–[3] typically use edge maps directly extracted from the photos as a substitute for real sketches during training, which are referred to as edge-based models. These methods have achieved unanimous success in edge-to-image translation. However, due to the huge structural discrepancy between the sketches and edge maps, they cannot adapt to sketch inputs and often generate very poor images (*e.g.*, Fig. 10(b)-(d)), which greatly restricts their application in practice. To solve this issue, some works study edge pre-processing [2] or collect real sketch datasets [4], [5]

Manuscript received March 9, 2021; revised August 27, 2021; accepted September 26, 2021. Date of publication October 21, 2021; date of current version October 28, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102702, in part by the Fundamental Research Funds for the Central Universities, and in part by the National Natural Science Foundation of China under Contract 62172020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiwen Lu. *(Corresponding author: Jiaying Liu.)*

Shuai Yang, Jiaying Liu, and Zongming Guo are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100080, China (e-mail: williamyang@pku.edu.cn; liujiaying@pku.edu.cn; guozongming@pku.edu.cn).

Zhangyang Wang is with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: atlaswang@utexas.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2021.3120669, provided by the authors.

Digital Object Identifier 10.1109/TIP.2021.3120669

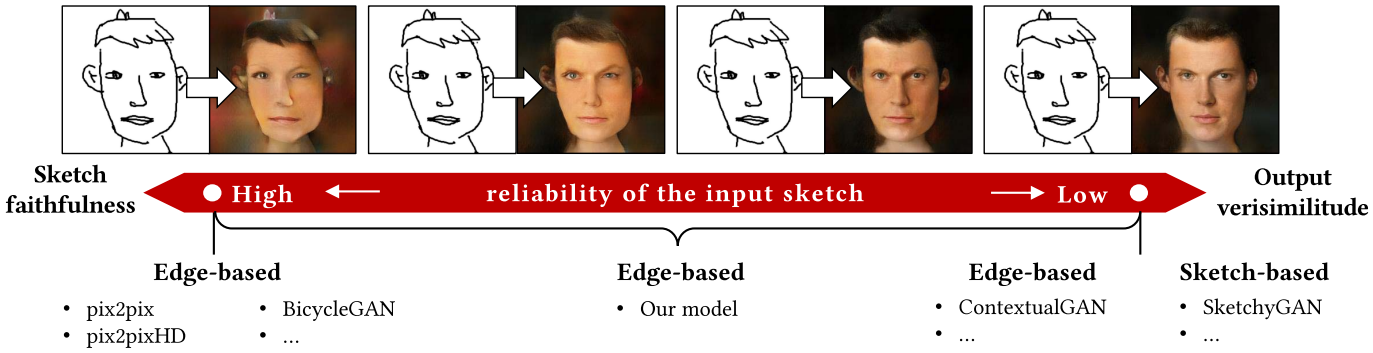


Fig. 2. Illustration of the sketch-to-image translation spectrum. Our model differs from existing models in that we allow users to define how "reliable" the input sketch should be considered for the final output, thus balancing between sketch faithfulness and output realism, which has not been well studied in previous approaches. In addition, as edge-based models, ContexualGAN [8] and our model realize realism without real sketch data for training, and our model further achieves controllability and efficiency.

to train sketch-based models [6], [7]. However, sketches of different users often have significantly varied appearances, far beyond the scope of the existing datasets, demanding higher generalizability and robustness of the model. Therefore, it has a high value to investigate adapting edge-based models to hand-drawn sketches.

Recently, ContextualGAN [8] provides an intuitive solution. It learns a joint edge-image manifold and searches the nearest neighbor to the input sketch within the manifold according to its edge part. Then, the accompanying retrieved image part has both photo-realistic properties and structural consistency with the sketch. Unlike the standard edge-based model that strictly hinges on the input sketch, ContextualGAN only uses the input as a weak reference as illustrated in Fig. 2. However, neither model provides users with the controllability on the sketch faithfulness, *i.e.*, to what extent we should stick to the given sketch? Failing to consider such controllability, the searched result by ContextualGAN can sometimes be too far from the input as we will show in Fig. 11, leaving little room for users to calibrate between freedom of sketching and the overall image realism: a key desirable feature for interactive photo synthesis.

Given all this, we are motivated to study a new problem of controllable sketch-to-image translation that works robustly on various hand-drawn sketches without the need for collecting real sketch training data. The key idea is to refine the sketches to match the fine structures of the edge maps and to provide users with a control parameter to freely adjust the refinement level to realize the aforementioned controllability. Fig. 1(a) intuitively displays our feature: supporting users to navigate across different refinement levels and choose the most ideal one to adapt a single model to extremely diverse sketches. There are two challenges to this problem. First, without real sketches, it is quite impossible to directly learn the mapping between sketches and edge maps. Second, for controllability, we must build more complicated multi-level mapping, which makes the problem more difficult.

In this paper, we present a novel robust and controllable sketch-to-image translation framework to meet these challenges. As inspired by the coarse-to-fine painting process of artists, we model the rough sketches as a drawable region covering the fine edges to specify where the final lines should lie. Then, the approximate mapping between sketches and edge maps can be well established by learning a translation from the drawable region, created by edge dilation, to the original edge maps. The coarse level of the sketches can be naturally controlled and adjusted by modifying the dilation radius. With rough sketches at various coarse levels, the multi-level mapping is built. Finally, we propose to leverage scale-aware style transfer to model the style features of coarse-level sketches and remove such dilation-based styles to obtain the refined output. Our model only uses color images and their extracted edge maps for training and can work robustly on diverse real-world sketches during testing. It can also serve as a plug-in for other edge-based models, providing refinement for their inputs to boost the performance. Fig. 1(a) shows the overall performance of our model on varied sketches.

Compared with our previous work [9], we further explore more advanced user controllability in terms of facial attribute editing, and spatially non-uniform refinement. First, we expand our model with facial attribute control, which supplements the controllability over features that cannot be captured by sketch images such as the hair colors as shown in Fig. 1(b). Then, we introduce a mask into our framework to specify the editing region as in Fig. 1(c). Moreover, we extend our model to spatially non-uniform refinement by expanding the scalar refinement level to a refinement level map and re-design the corresponding dilation operation and style transfer operation. It enables users to preserve important structural details while refining the raw sketches, as shown in Fig. 1(d). In addition, thorough experiments are conducted to clarify the effectiveness of the proposed method, including additional comparison results for both quantitative and qualitative evaluation, and results for facial attribute control and spatially non-uniform refinement level control. Our contributions are summarized as four-folds:

- We explore a new problem of controllable sketch-to-image translation for face image synthesis, which adapts edge-based models to real-world hand-drawn sketches, where the users have the freedom to balance the sketch faithfulness with the output realism.
- We propose a sketch refinement method using coarse-to-fine dilations, inspired by the painting process of artists,

which bridges the gap between coarse-level sketches and fine-level edges.
- We propose a style-based network architecture, which successfully learns to refine the input sketches into diverse and continuous levels.
- We present advanced user controllability with respect to facial attribute editing, editing region control and spatially non-uniform refinement.

The rest of this paper is organized as follows. Section II reviews related works in image-to-image translation, sketch-based image synthesis and image inpainting. In Section III, we elaborate on the key idea of dilation-based sketch refinement and the proposed controllable sketch-to-image translation framework. Section IV presents how we incorporate new features of editing region control, facial attribute control and the spatially non-uniform sketch refinement into the proposed framework. In Section V, we verify the superiority of the proposed method through comprehensive experiments and comparisons with the state-of-the-art image-to-image translation methods. Finally, we conclude our work in Section VI.

II. Related Work

A. Image-to-Image Translation

The goal of image-to-image translation is to transform an image between a source domain and a target domain. It is first raised by Isola *et al.* [1], where a powerful framework pix2pix is designed to learn mappings between paired data. Subsequent researches improve pix2pix [1] in terms of the increase in image resolution [10], multi-modal translation [11] and multi-domain translation [12], [13]. In [14], spatially-adaptive normalization is proposed to better inject the conditional image information in a multi-scale manner. Later, Zhu *et al.* [15] put forward improved semantic region-adaptive normalization to introduce styles in a local manner for semantic face editing. Another important research direction lies in learning mappings between unpaired data, where the pioneering work of CycleGAN [16] presents a cycle consistency constraint to provide practical pixel-level loss. Based on this idea, UNIT [17] and MUNIT [18] assume a shared latent space across different domains, which diversify the generated images. A contextual loss [19] is proposed to provide more robust feature-level cycle consistency.

B. Sketch-to-Image Translation

Edge-based sketch-to-image translation methods [1], [20], [21] use edge maps to train the model to avoid the tedious collection of sketches. These methods can be further adapted to image editing tasks by introducing a mask to specify the editing region of a given photo [2], [3], [22]. The main issue for the edge-based models is that they are hard to generalize to real-world sketches, which have significant structural discrepancy from the edge maps. Although some sketch datasets [4], [5] have been collected for the training of sketch-based models [6], [7], current datasets are far from meeting the actual diversified research needs.

To solve this issue, Lu *et al.* [8] propose ContextualGAN to search the nearest neighbor of the given sketch in an edge manifold learned by GANs, which avoids exactly following the sketch and generates plausible results. Our method shares the same goal as ContextualGAN to adapt edge-based models to sketch inputs but further takes the controllability into account, allowing users to select the degree of refinement. Moreover, the iterative optimization-based nearest neighbor search process is computationally expensive. By comparison, the proposed feed-forward model provides efficient, flexible and user-friendly face synthesis and editing tools.

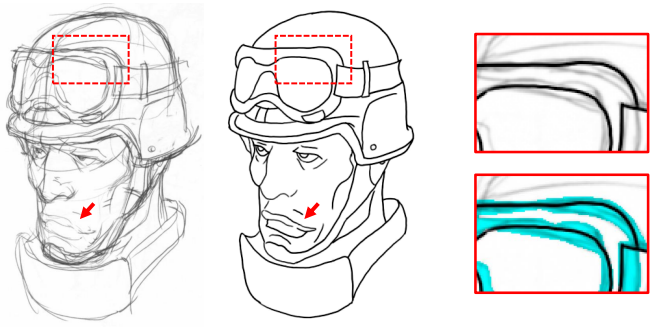


Fig. 3. Rough sketch (left) to fine sketch (middle). Lines in the red boxes are enlarged and overlayed on the right. Rough sketches form a cyan drawable region to indicate where the fine sketches should lie. The red arrow points to the new structures inferred from the sketch. Copyright: Krenz Cushart [33].

C. Image Inpainting

Image inpainting investigates the content reconstruction within a region of an image specified by a mask. It provides a potent tool for image editing such as removing unwanted objects and modifying image structures or details. Traditional exemplar-based models [23]–[27] fill the masked region with the pixels in the known region, which cannot synthesize unseen content. Recently, data-driven models such as Context Encoder [28] and DeepFill [29], [30] leverage large-scale data to train inpainting networks, which achieves intelligent semantic-aware completion. Follow-ups improve Context Encoder in terms of high-resolution image inpainting [31], free-form masks [30], and output diversification [32].

III. Controllable Sketch-to-Image Translation

The goal of our controllable sketch-to-image translation problem is to design and train a novel sketch refinement network G without using sketch data. G revises the inaccurate structures of the hand-drawn sketches to match those of the edge maps so that the refined sketches can be smoothly fed into existing edge-to-image translation models F to generate photo-realistic images. In other words, G aims to adapt F to the hand-drawn sketches. To take a step further, we condition G by a key parameter $\ell \in [0, 1]$ to control the refinement level: Users can apply stronger refinement by increasing ℓ.

A. Sketch Refinement via Dilation

Our key idea is to model the rough sketch as a dilated drawable region and to shrink and refine it to get fine sketches, following the coarse-to-fine painting process of human artists.

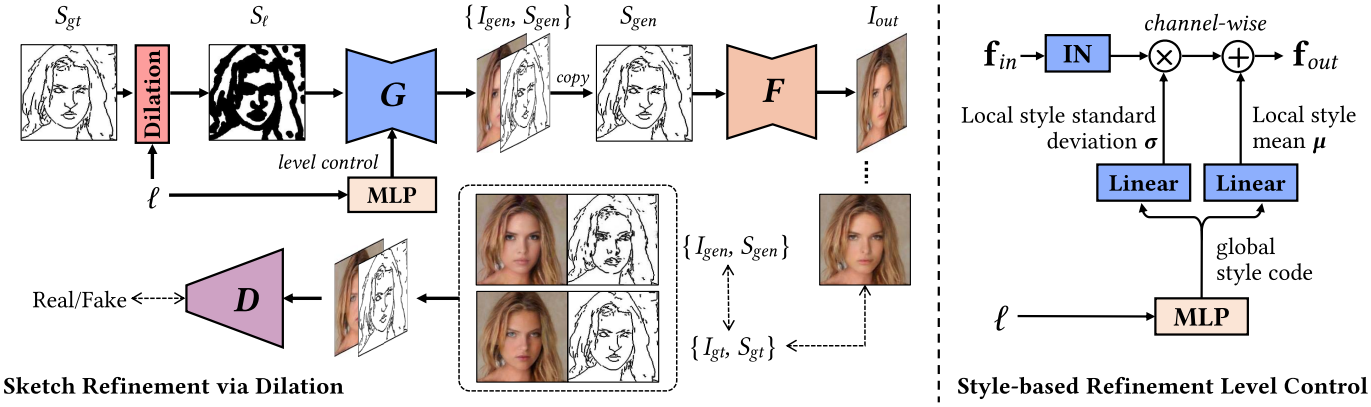


Fig. 4. Framework overview. A novel sketch refinement network G is proposed to refine the rough sketch S_ℓ modelled as a dilated drawable region to match the fine edge S_{gt}. The refined output S_{gen} is fed into a pretrained edge-based model F to obtain the final result I_{out}. A parameter ℓ is introduced to control the refinement level. It is realized by encoding ℓ into the style codes and performing a style-based adjustment over the feature map $\mathbf{f}_{in}$ of the convolutional layer of G to remove the dilation-based styles.

Fig. 5. Rough sketch generation. (a) S_{gt}. (b) Deformed edges with lines discarded. (c) $\Omega(S_{gt})$. (d) Overlay red S_{gt} above $\Omega(S_{gt})$ with discarded lines tinted in cyan.

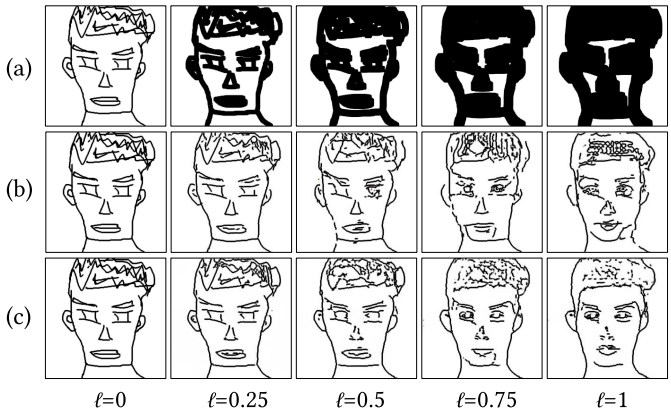


Fig. 6. Sketch refinement at different ℓ. (a) Dilated S_{gt} using different r. (b) Refinement results by pix2pix [1] trained separately for each r. (c) Refinement results by our proposed single model with multi-level control.

Fig. 3 shows an example of a rough sketch, which contains redundant lines. These lines are usually drawn at the beginning of the painting to roughly determine the position and shape of the object. Then artists gradually merge lines, add details and fix errors to refine the sketches. If we overlay the fine sketch on the rough sketch, we can find that the redundant lines form a drawable region (cyan region in Fig. 3) covering the final lines. Thus, essentially, the coarse-to-fine painting process is to progressively shrink and refine the drawable region.

Based on the observation, sketch refinement can be naturally modelled as an image-to-image translation problem between rough sketches and fine sketches, as illustrated in Fig. 4. For our edge-based model, we use the edge map S_{gt} extracted from the face image I_{gt} as the fine sketch, while its rough counterpart is defined as a drawable region $\Omega(S_{gt})$ (denoted as S_ℓ in Fig. 4) covering S_{gt} with Ω the proposed novel dilation operation. Next, we will introduce the proposed dilation-based drawable region generation algorithm, which automatically synthesizes $\Omega(S_{gt})$ from S_{gt} to build our training data.

Rough Sketch Data Generation: Fig. 5 illustrates the pipeline of the proposed drawable region generation. The key idea is to use a dilation operator in mathematical morphology for expanding distorted and incomplete edges into drawable regions. Specifically, we first randomly deform the fine edges to simulate the structural inaccuracies in hand-drawn sketches. To ensure that each ground truth edge is completed covered by its distorted and dilated result, the offset of each pixel after deformation is limited to less than r, with r the radius of dilation. Furthermore, as pointed by the red arrow of Fig. 3, artists will also infer and add new structures during refinement. To this end, we further discard partial lines by removing random patches from the full sketches. Therefore, our network is motivated to learn to complete the missing structures (*e.g.*, cyan lines in Fig. 5(d)). Finally, dilation is applied to generate sufficient training data $\{\Omega(S_{gt}), S_{gt}\}$. Note that deformation and discard are only applied during the training phase.

Intuitively, wide drawable regions provide more room for refinement, corresponding to a higher refinement level. In Fig. 6(b), we verify this statement via a simple experiment, where for each dilation radius we train a pix2pix [1] to map $\Omega(S_{gt})$ to S_{gt}. The trained model effectively refines an unseen hand-drawn sketch, and as the radius grows, the degree of refinement also increases. This property suggests a good potential for convenient refinement level control. The next section introduces how we integrate multi-level refinement into a single model, which achieves both practical controllability and more robust performance, as illustrated in Fig. 6(c).

B. Controllable Sketch-to-Image Translation

In our sketch-to-image translation task, we have a hand-drawn face sketch image S to serve as shape guidance for the model to infer a corresponding photo-realistic face image I. Users adjust the refinement level determined by a parameter ℓ, where a larger ℓ will make the model more intensely revise the lines in S to generate realistic facial structures.

Our training requires no hand-drawn sketches. Instead, we use edge maps S_{gt} of the real-world face photos I_{gt} and their corresponding drawable region $\Omega(S_{gt})$ to train our model. As analyzed in Sec. III-A, the refinement level is positively correlated with the dilation radius r. Thus, to incorporate level control, we introduce ℓ in the drawable region generation as $S_\ell = \Omega_\ell(S_{gt})$ by using ℓ to determine the value of r, where $r = \ell R$ and R is the maximum allowable radius. Then, our task is simply to train G to map S_ℓ to S_{gt} with conditional ℓ, as illustrated in Fig. 4. Specifically, G accepts S_ℓ, with middle layers adjusted by ℓ, and produces a four-channel tensor: the RGB image I_{gen} and the refined one-channel sketch S_{gen}, *i.e.*, $(I_{gen}, S_{gen}) = G(S_\ell, \ell)$. Here, the additional output I_{gen} provides perceptual guidance for the edge generation and serves as the final image output if F is unavailable, meaning that G can be used independently. Finally, a discriminator D is used to improve the results via adversarial training.

1) Style-Based Refinement Level Control: To learn robust mappings from S_ℓ at different coarse levels to S_{gt} using a single G, our key idea is to regard S_ℓ as a stylish version of S_{gt} with a style associated with the proposed dilation operation. Sketches at different coarse levels have different styles. Then our task becomes to destylize S_ℓ back to S_{gt}. Inspired by AdaIN-based style transfer [34] and image generation [35], we propose an effective yet simple style-based module to accomplish it. AdaIN [34] models the style as style parameters of the mean μ and standard deviation σ of the feature $\mathbf{f}$ and transfers the style via distribution scaling and shifting:

$$\mathbf{f}' = \text{AdaIN}(\mathbf{f}, \mathbf{f}_s) = \sigma(\mathbf{f}_s)\left(\frac{\mathbf{f} - \mu(\mathbf{f})}{\sigma(\mathbf{f})}\right) + \mu(\mathbf{f}_s), \tag{1}$$

where $\mathbf{f}$ is first instancely normalized using its channel-wise mean $\mu(\mathbf{f})$ and standard deviation $\sigma(\mathbf{f})$, and then denormalized by matching these style parameters to those of the style reference $\mathbf{f}_s$. Note that the same operation can also be used for its reverse process, *i.e.*, destylization. Specifically, to obtain the original $\mathbf{f}$, $\mathbf{f}'$ should also be first normalized and then denormalized to match the style of $\mathbf{f}$. In our problem, the style parameters are related to the condition ℓ. Therefore, as shown in Fig. 4, ℓ is first encoded into a global style code via a multi-layer perceptron, which is then mapped to local style mean μ_ℓ and variance σ_ℓ via two affiliated linear layers within each middle convolution layer of G for AdaIN-based destylization:

$$\mathbf{f}_{out} = \sigma_\ell\left(\frac{\mathbf{f}_{in} - \mu(\mathbf{f}_{in})}{\sigma(\mathbf{f}_{in})}\right) + \mu_\ell, \tag{2}$$

where $\mathbf{f}_{in}$ and $\mathbf{f}_{out}$ are input and output features, respectively.

2) Loss Function: G is tasked to approach the ground truth photo and sketch:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{I_{gt},\ell}[\|I_{gen} - I_{gt}\|_1 + \|S_{gen} - S_{gt}\|_1 + \|I_{out} - I_{gt}\|_1], \tag{3}$$

where $I_{out} = F(S_{gen})$ is the final output in our problem. By additionally consider the quality of I_{out}, G can be well adapted to F in an end-to-end manner. Perceptual loss $\mathcal{L}_{\text{perc}}$ [36] to measure the semantical similarity of photos is further used:

$$\mathcal{L}_{\text{perc}} = \mathbb{E}_{I_{gt},\ell}[\sum_i \lambda_i(\|\Phi_i(I_{gen}) - \Phi_i(I_{gt})\|_2^2 + \|\Phi_i(I_{out}) - \Phi_i(I_{gt})\|_2^2)], \tag{4}$$

where $\Phi_i(x)$ is the feature of x in the i-th layer of VGG19 [37] and λ_i is the layer weight. Finally, we use hinge loss as our adversarial objective function:

$$\mathcal{L}_G = -\mathbb{E}_{I_{gt},\ell}[D(I_{gen}, S_{gen})], \tag{5}$$

$$\mathcal{L}_D = \mathbb{E}_{I_{gt},\ell}[\text{ReLU}(\tau + D(I_{gen}, S_{gen}))] + \mathbb{E}_{I_{gt}}[\text{ReLU}(\tau - D(I_{gt}, S_{gt}))], \tag{6}$$

where τ is a margin parameter.

IV. Attribute and Spatial Control

A. Editing Region Control for Sketch-Based Image Editing

In this section, we introduce a mask into our framework to specify the editing region. As shown in Fig. 1(c), it allows users to further refine the local regions of a synthesized image, or to edit the specified region of a given real-world photo, such as adding sunglasses and removing bangs.

In the image editing task, a mask M to indicate the editing region and an image $I_{in} = I_{gt} \odot (\mathbf{1} - M)$ to be edited are given as additional inputs, where $\odot$ is the element-wise multiplication operator. A hand-drawn sketch S serves as soft structural guidance in the masked region, which will be refined by the model according to the refinement level ℓ.

Similar to the sketch-to-image translation task, a masked drawable region $S_\ell = \Omega_\ell(S_{gt}) \odot M$ is used for training with no real sketches required. G is tasked to map S_ℓ back to S_{gt} based on the contextual condition I_{in}, the spatial condition M and the level condition ℓ. As illustrated in Fig. 7, I_{in}, S_ℓ and M are concatenated and fed into G, which outputs I_{gen} and S_{gen}, *i.e.*, $(I_{gen}, S_{gen}) = G(I_{in}, S_\ell, M, \ell)$. Finally, F yields the final image output I_{out} based on I_{in}, S_{gen} and M.

In terms of the loss function, $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{perc}}$ take the same form as in Sec. III-B. The only adaptation is that D takes M as an additional input to pay more attention to the realism of the masked region:

$$\mathcal{L}_G = -\mathbb{E}_{I_{gt},M,\ell}[D(I_{gen}, S_{gen}, M)], \tag{7}$$

$$\mathcal{L}_D = \mathbb{E}_{I_{gt},M,\ell}[\text{ReLU}(\tau + D(I_{gen}, S_{gen}, M))] + \mathbb{E}_{I_{gt},M}[\text{ReLU}(\tau - D(I_{gt}, S_{gt}, M))]. \tag{8}$$

B. Attribute Control

Although sketches are highly expressive, they cannot provide information other than structural cues, such as the important color cues. Some sketches can even be too abstract to provide accurate structural cues, such as the gender in the poor sketch in Fig. 1(a). To tackle this problem, in this section, we expand our framework with facial attribute control.

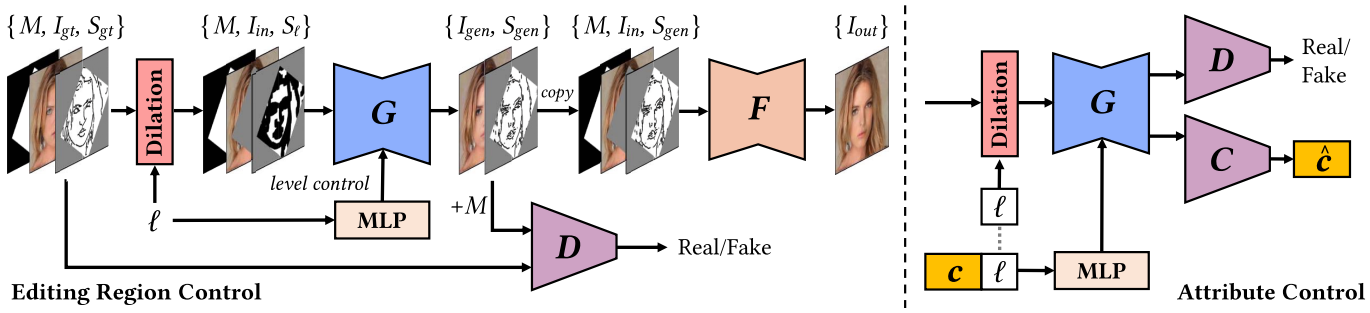


Fig. 7. Illustration of editing region control and facial attribute control. Editing region control: An additional mask M is provided to specify the area to modify. G is fed with M and the masked photo I_{in} to infer the information in the masked area. D learns to distinguish between real and fake images/sketches in the masked area. Facial attribute control: Our model allows users to control the facial attribute via a class label c. c is concatenated to ℓ and is encoded into the style code to adjust the output of G. An additional classifier C learns to minimize the classification error for the label c.

Our key idea is to regard facial attributes as a kind of face style and naturally integrate them into our proposed style-based refinement level control method. Specifically, we first extract facial attributes from I_{gt} as a one-hot vector c, where each element corresponds to one attribute with 1 indicating the image satisfies this attribute and 0 otherwise. As shown in Fig. 7, the concatenation of c and ℓ is then fed into the multi-layer perceptron to obtain the global style code for the AdaIN-based style transfer. Finally, we employ an auxiliary classifier C [38] to predict the facial attribute of the generated images, which should approach the input c. Therefore, apart from the original losses, a new cross-entropy loss is proposed:

$$\mathcal{L}_{cls} = \mathbb{E}_{I_{gt},\ell,c}[-c^{\intercal}\log[C(I_{gt})] - c^{\intercal}\log[C(I_{gen})]], \quad (9)$$

where I_{gen} is the RGB image output of $G(S_\ell, \ell, c)$ and $c^{\intercal}$ is the transposed c. Intuitively, C is trained to predict the ground truth c from I_{gt}, and G tries to generate I_{gen} with the correct attributes to make C give the right prediction.

C. Spatially Non-Uniform Refinement

The proposed method provides a global refinement over the input sketches. However, the accuracy of the structure can vary within one sketch, which demands spatially non-uniform sketch refinement for more flexible controllability. Our model can be easily extended to meet this demand. We expand the scalar refinement level ℓ to a refinement level map L, which has the same resolution as the sketches. Then the spatially non-uniform dilation operation and AdaIN operation are designed to adapt to L, which can be directly applied during testing without retraining the network.

To be specific, as shown in Fig. 8, we first enumerate ℓ from a set $N_\ell = \{k/R | k = 0, 1, \ldots, R\}$ where R is the maximum allowable dilation radius. Then, for each $\ell \in N_\ell$, its weighting map W_ℓ is calculated as

$$W_\ell = \max(1 - |\ell - L|R, 0), \quad (10)$$

which selects the regions corresponding to the range of $[\ell - 1/R, \ell + 1/R]$ in L. Finally, the spatially non-uniform dilation operation is defined as:

$$S_L = \sum_{\ell \in N_\ell} W_\ell S_\ell. \quad (11)$$

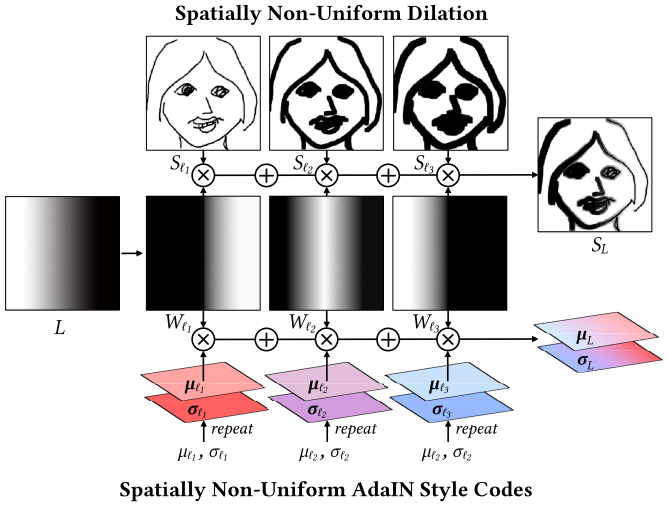


Fig. 8. Illustration of spatially non-uniform refinement level control. ℓ is enumerated (in this simple case, we use three levels $\ell_1 < \ell_2 < \ell_3$). The dilated sketches and style codes under ℓ_1, ℓ_2, ℓ_3 are fused by the weighted average for spatially non-uniform dilation and adaptive instance normalization, where the weights Ws are derived from the level map L.

Likewise, for spatially non-uniform AdaIN operation, the style parameters $\{\mu_\ell, \sigma_\ell\}$ corresponding to ℓ are expanded to the same spatial resolution as the feature map $\mathbf{f}_{in}$. We denote them as $\{\boldsymbol{\mu}_\ell, \boldsymbol{\sigma}_\ell\}$. And the final style parameters for L are

$$\boldsymbol{\mu}_L = \sum_{\ell \in N_\ell} W_\ell \boldsymbol{\mu}_\ell, \quad \boldsymbol{\sigma}_L = \sum_{\ell \in N_\ell} W_\ell \boldsymbol{\sigma}_\ell, \quad (12)$$

which are used for style transfer (Eq. (2)). Note that L is always resized to match the resolution of $\mathbf{f}_{in}$.

Fig. 9 presents a toy example of the proposed spatially non-uniform refinement, where a set of $\hat{L}$ with horizontally changing refinement levels is given. Our method successfully optimizes the facial structures in the half designated area, while leaving the other half area less refined. Note that our previous conference model [9] can only handle the cases on the diagonal (when $\ell_{min} = \ell_{max}$). The proposed extension provides more flexible controllability.

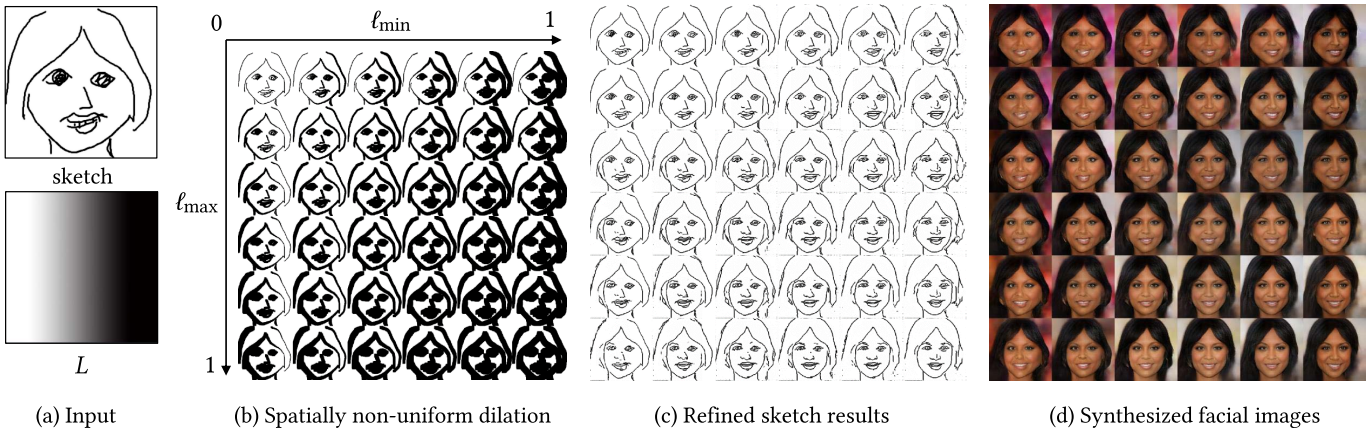


Fig. 9. Results of spatially non-uniform refinement level control. (a) Input sketch and level map. (b) Spatially non-uniform dilation results using the level map $\hat{L} = (\ell_{\max} - \ell_{\min})L + \ell_{\min}$. (c) Sketch refinement results using $\hat{L}$. (d) Facial synthesis results using $\hat{L}$.

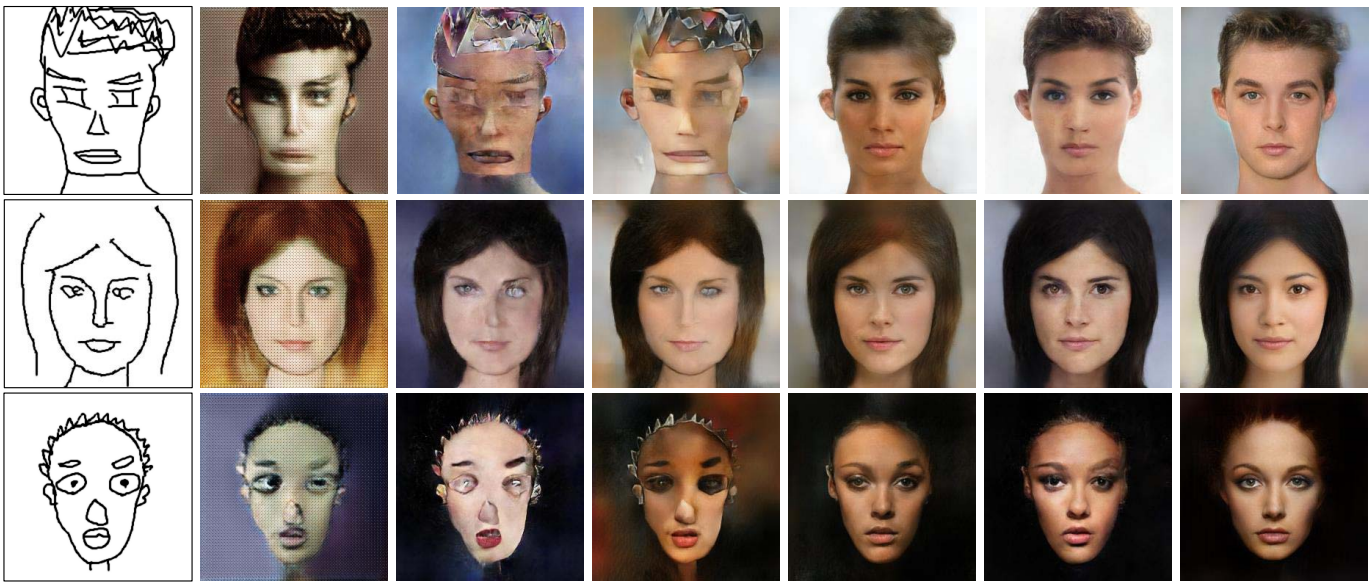


Fig. 10. Comparison with state-of-the-art methods on facial synthesis. (a) Input hand-drawn sketches. (b) BicycleGAN [11]. (c) pix2pixHD [10]. (d) pix2pix [1]. (e) Our results with $\ell = 1$. (f) pix2pixHD using our refined sketches as input. (g) Our results with facial attribute control and spatially non-uniform refinement level control. Note that results in (g) requires additional attribute supervision. We only use it to demonstrate the improved controllability over (e).

V. Experimental Results

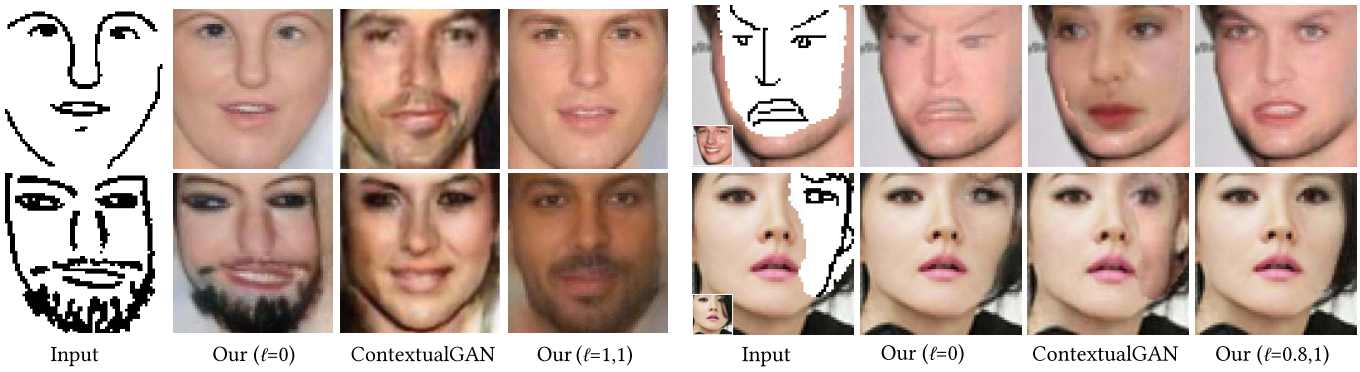
A. Implementation Details

1) Dataset: We use CelebA-HQ dataset [39] with edge maps extracted by HED edge detector [40] to train our model. The masks are generated as randomly rotated rectangular regions following [2]. To make a fair comparison with ContextualGAN [8], we also train our model on CelebA dataset [41].

2) Rough Sketch Generation: We implement dilation operations as convolutional layers with all-ones kernels of different radii r, followed by data clipping into the range [0, 1]. Dilation results using the fractional radii are obtained by interpolating the results under the integer radii. We use the sampling layer [42] to deform lines. We generate random offset maps via Gaussian noises and resample the input sketch based on the offset maps via the sampling layer. We randomly erase $0 \sim 3$ rectangular regions with height and width in $[8, 24]$ to discard lines, which can improve the robustness.

3) Network Architecture: The generator G takes the Encoder-ResBlocks-Decoder architecture [36] with skip connections [1] to preserve the low-level information. Each convolutional layer is followed by an AdaIN layer [34] except the first and the last layer. The discriminator D follows the SN-PatchGAN [30] to stabilize training. The classifier C shares layers with D expect the last layer. Finally, we use pix2pix [1] as our edge-based baseline model F.

4) Network Training: Our network is first trained with $\ell = 1$ for 30 epoches, and then trained with uniformly sampled $\ell \in [0, 1]$ for 200 epoches. The maximum allowable dilation radius R is 10 and 4 for CelebA-HQ [39] and CelebA [41],



(a) Comparison on facial synthesis (b) Comparison on facial editing

Fig. 11. Comparison with ContextualGAN [8] on facial synthesis and facial editing.

respectively. For all experiments, the weight for $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{perc}}$, $\mathcal{L}_G$, $\mathcal{L}_{\text{D}}$ and $\mathcal{L}_{\text{cls}}$ are 100, 1, 1, 1 and 10, respectively. Conv2_1 and conv3_1 layers of the VGG19 [37] weighted by 1 and 0.5 are used to compute $\mathcal{L}_{\text{perc}}$. For hinge loss, we set τ to 10.

B. Comparisons With State-of-the-Art Methods

1) Face Synthesis: Fig. 10 presents the visual comparison on face synthesis with two state-of-the-art image-to-image translation models: BicycleGAN [11] and pix2pixHD [10]. The two models and the baseline model F (pix2pix [1]) are trained on edge images and strictly follow the distorted sketch inputs. Therefore, their results are poor and unrealistic. By comparison, the proposed method only takes the sketch as useful yet flexible constraints, and successfully revises the inaccurate facial structures, striking a good balance between realism and consistency with the sketch inputs. To verify the effectiveness of our sketch-edge input adaption, we directly use our refined sketches as the inputs of pix2pix-HD without fine-tuning on it, and the visual quality of its results is significantly enhanced as shown in Fig. 10(f). Finally, we include the results of our extended model with facial attribute control and non-uniform refinement level control in Fig. 10(g). Such extra condition information eases the training and inferencing of the model, leading to more realistic results with less artifacts.

ContextualGAN [8] is the most related model to ours that treats sketch inputs as soft guidance. Fig. 11(a) further presents a comparison with it on CelebA dataset. Although ContextualGAN synthesizes realistic faces, it fails to preserve important facial attributes provided by user sketches such as the beard. The reason is that its learned edge-image manifold may collapse for infrequent attributes and at the same time, the search result can sometimes travel too far from the input during optimization. By comparison, our model achieves better structural consistency. Moreover, ContextualGAN requires costly iterative optimization for the nearest neighbor search, which is less efficient than the proposed feed-forward method. Specifically, our implemented ContextualGAN requires about 7.89 s per 64×64 image with a GeForce GTX 1080 Ti GPU, while our model only takes about 12 ms per image.

TABLE I
USER PREFERENCE RATIO ON FACE SYNTHESIS

Dataset	CelebA-HQ			CelebA	
Method	BicycleGAN	pix2pixHD	Ours	ContextualGAN	Ours
Score	0.024	0.031	**0.945**	0.094	**0.906**

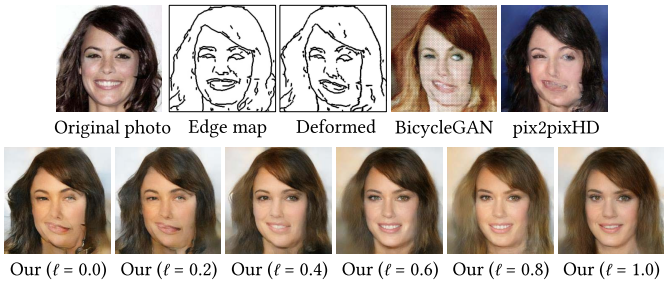


Fig. 12. Visual comparison with state-of-the-art methods on simulated poor CelebA-HQ sketches. Top row, from left to right: The original CelebA-HQ facial image and its edge map, the simulated poor sketch by deforming the edge map, facial synthesis results by BicycleGAN [11] and pix2pixHD [10]. Bottom row, from left to right, our facial synthesis results with $\ell = 0.0$, 0.2, 0.4, 0.6, 0.8, 1.0, respectively.

To better understand the performance of the compared methods, we conduct quantitative evaluations on both hand-drawn sketches and synthetic data. First, a user study is conducted on real data, where we collected a total of 38 hand-drawn sketches and asked participants to choose the best result to balance the output realism and the sketch faithfulness. Finally, a total of 20 subjects participate in this study and a total of 760 selections are tallied. The preference ratio is used as the evaluation metrics. It measures the percentage of times a method is selected in all its related selections. According to the definition, if a method performs significantly better than all other methods, its mean preference ratio can reach 1.0. Table I demonstrates the preference scores, where the proposed method receives notable preference for both sketch detail preservation and output naturalness.

In addition to the real data, we also generated 100 distorted CelebA-HQ edge maps as pseudo sketches for quantitative evaluation. Table II reports the perceptual loss and Fréchet

TABLE II
RECONSTRUCTION QUALITY OF THE COMPARISON METHODS ON SIMULATED POOR CELEBA-HQ SKETCHES

Method	BicycleGAN	pix2pixHD	pix2pix (Ours with $\ell = 0$)	Ours				
				$\ell = 0.2$	$\ell = 0.4$	$\ell = 0.6$	$\ell = 0.8$	$\ell = 1$
$\mathcal{L}_{\text{perc}}$	447.617	218.084	201.676	191.026	183.753	**179.659**	180.293	180.557
FID	152.857	117.968	121.050	115.658	**108.455**	109.117	113.132	113.669

Fig. 13. Comparison with state-of-the-art methods on facial editing. (a) Input photos, masks and sketches. (b) DeepFillv2 [30]. (c) SC-FEGAN [3]. (d) Our results with $\ell = 0$. (e) Our results with $\ell = 1$. (f) SC-FEGAN using our refined sketches as input.

inception distance (FID) [43] between the synthesized images and the original photos, which measure how well a method revises the inaccurate structures and the output realism, respectively. Our method achieves the best scores in both metrics. Moreover, an evident improvement is observed as the refinement level ℓ starts to increase. Then, very large dilation radii will eliminate important details, thus the performance slightly drops when $\ell > 0.6$. The visual result in Fig. 12 confirms this conclusion. Our method is quite robust to flawed structures than the compared methods. Excessive refinement will remove some details in the original image, such as the degree of smile.

2) Face Editing: In Sec. IV-A, we extend our method to face editing. Fig. 13 presents the visual comparison on face editing with two state-of-the-art inpainting models: DeepFillv2 [30] and SC-FEGAN [3]. The released DeepFillv2 uses no sketch guidance, which means the reliability of the input sketch is set to zero ($\ell = \infty$). Despite being one of the most advanced image inpainting models, DeepFillv2 fails to restore facial structures well, indicating the necessity of user guidance. Meanwhile, SC-FEGAN is a representative face editing model with edge guidance. However, it strictly synthesizes facial structures following rough sketches and generates unrealistic details. Similarly, the poor results of the base model F in Fig. 13(d) also suggest the need for sketch refinement. By comparison, our method successfully revises the inaccurate structures and generates more natural face details. Finally, as in the face synthesis task, we also test our refined sketches on the other edge-based model SC-FEGAN, and observe an obvious improvement without any fine-tuning.

We have also adapted ContextualGAN to the face editing task through additionally considering the similarity with the known region of the photo when searching the nearest neighbor. Fig. 11(b) compares the proposed method with ContextualGAN. It can be seen that ContextualGAN generates abrupt inpainting boundaries. The reason might be that its learned manifold fails to cover the real facial distribution and thus it cannot search for a good solution to match the known region. As a comparison, the proposed method synthesizes more natural faces.

C. Ablation Study

To analyze the effect of each module of the proposed method on sketch refinement, we perform ablation studies in this section.

1) Rough Sketch Modelling: First of all, we study our key dilation-based rough sketch generation, including three operations of line deformation, discard and dilation. Fig. 14 compares the effect of each operation. Dilation operation makes the network infer local facial details such as eyes, but it is not enough to force the network to revise structure errors. Line deformation helps the network learn to refine facial structures. Together with line discard, the network starts to repair missing structures. Finally, compared with learning a single-level refinement, training a single model on multi-level sketch refinement further enhances the overall performance, likely due to the fact that coarse-level refinement can benefit from the learned more robust fine-level features.

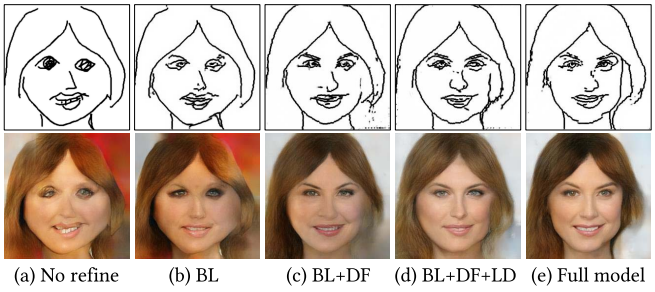


Fig. 14. Effect of rough sketch models. (a) Input sketch and generated image without refinement. (b)-(d) Refinement results using different rough sketch models. (b) Baseline: edge dilation with a fixed single dilation radius. (c) Baseline + line deformation. (d) Baseline + line deformation and discarding. (e) Edge dilation with multiple radii + line deformation and discarding.

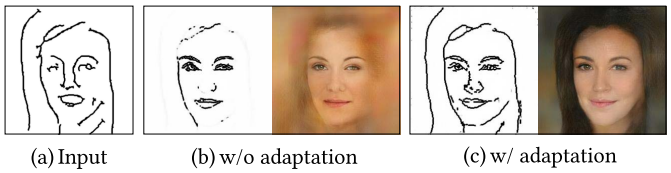


Fig. 15. Effect of adaptation to F.

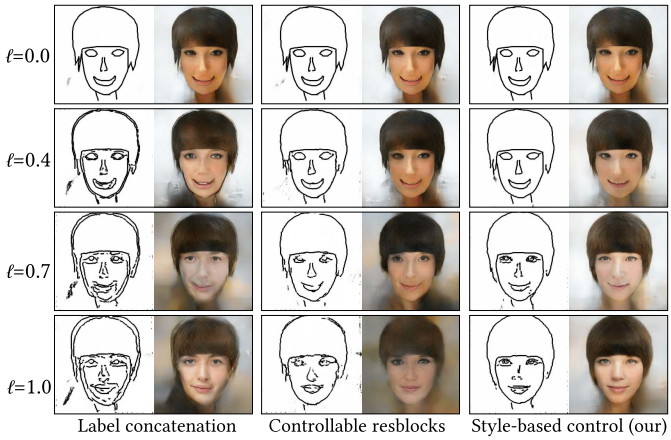


Fig. 16. Visual comparison on label conditioning strategy.

TABLE III
ACCURACY OF FACIAL ATTRIBUTE CLASSIFICATION WITHOUT AND WITH ATTRIBUTE CONTROL

Attribute	Hair color			Race			Aged			Gender		
ℓ	0	0.5	1	0	0.5	1	0	0.5	1	0	0.5	1
w/o AC	0.49	0.47	0.45	0.72	0.72	0.70	0.88	0.86	0.85	0.94	0.94	0.92
w/ AC	**0.90**	**0.96**	**0.99**	**0.92**	**0.99**	**1.00**	**0.94**	**0.98**	**1.00**	**0.99**	**1.00**	**1.00**

2) Adaptation to F: In Eqs. (3)(4), we adapt G to F by considering the reconstruction quality of the final output I_{out}. To examine the effect of such adaptation, we compare the results with and without the loss terms related to I_{out} in our loss function in Fig. 15. Without adaptation, the network is still able to refine the sketches, but the refined lines are obscure and light-colored, which cannot be recognized by F to generate the corresponding facial structures. It is likely because, without adaptation, the sketch output is only constrained by a pixel-level reconstruction loss, which is not robust since the black pixels only cover an extremely low proportion of the sketch image. The proposed adaptation serves as a kind of perceptual loss [36] for sketches, which guides G to yield lines that are fully perceivable by F. Therefore, our full model produces distinct sketches and clear facial photos.

3) Refinement Level Control: In Fig. 16, we present a comparison with three label conditioning strategies for refinement level control: label concatenation, controllable resblock [44] and the proposed style-based control. Label concatenation produces weird multiple lines for each edge. Controllable resblock yields clear structures but coarse facial details. Our style-based control is superior to the compared strategies for both natural structures and vivid facial details.

D. Fine-Grained Control

In Fig. 17, we present the new feature of fine-grained control provided by our extended model. Compared with our base model [9] in Fig. 17(b), by introducing facial attribute labels as additional guidance to relieve ambiguity, it becomes easier for our model to learn the challenging coarse-to-fine mapping. Therefore, the overall performance is enhanced as shown in Fig. 17(c). Our method generates realistic faces with cleaner facial structures. Furthermore, it enables users to adjust facial attributes to better match the input sketches. For example, without facial attribute labels, our model will be biased towards synthesizing female faces under large dilation due to the ambiguity and imbalanced distribution of the training data. This problem is well solved by manually setting the conditional attributes as shown in Fig. 17(d). Finally, our extended model allows users to apply spatially non-uniform refinement. In Fig. 17(e), our method largely corrects the inaccurate facial structures while preserving key features such as the raised corners of the mouth and eyebrows, which makes the synthesized faces more distinctive.

1) Quantitative Evaluation: To quantitatively evaluate the effectiveness of the facial attribute control, we report the accuracy of facial attribute classification. Specifically, we synthesize face images I_{gen} from the edge maps of 1,000 testing images in CelebA-HQ [39] with our base model and the extended model with attribute control (AC). For the extended model, the ground truth facial attributes are provided as inputs. The auxiliary classifier C is used to predict the facial attributes from I_{gen}, and the classification accuracies in terms of hair color, race, aged and gender under different refinement levels are reported in Table III. It can be seen that our extended model with AC outperforms the base model, and the advantage becomes more evident in attributes that cannot be sufficiently depicted by the sketches alone like hair color and race. Interestingly, as the refinement level increases, the classification accuracy of the base model drops since the dilated edges become less reliable. By comparison, the accuracy of our extend model improves because the model has more space to adjust the edges to match the target attributes.

Next, we investigate whether our extended model can adjust sketches spatially non-uniformly according to the label map L.

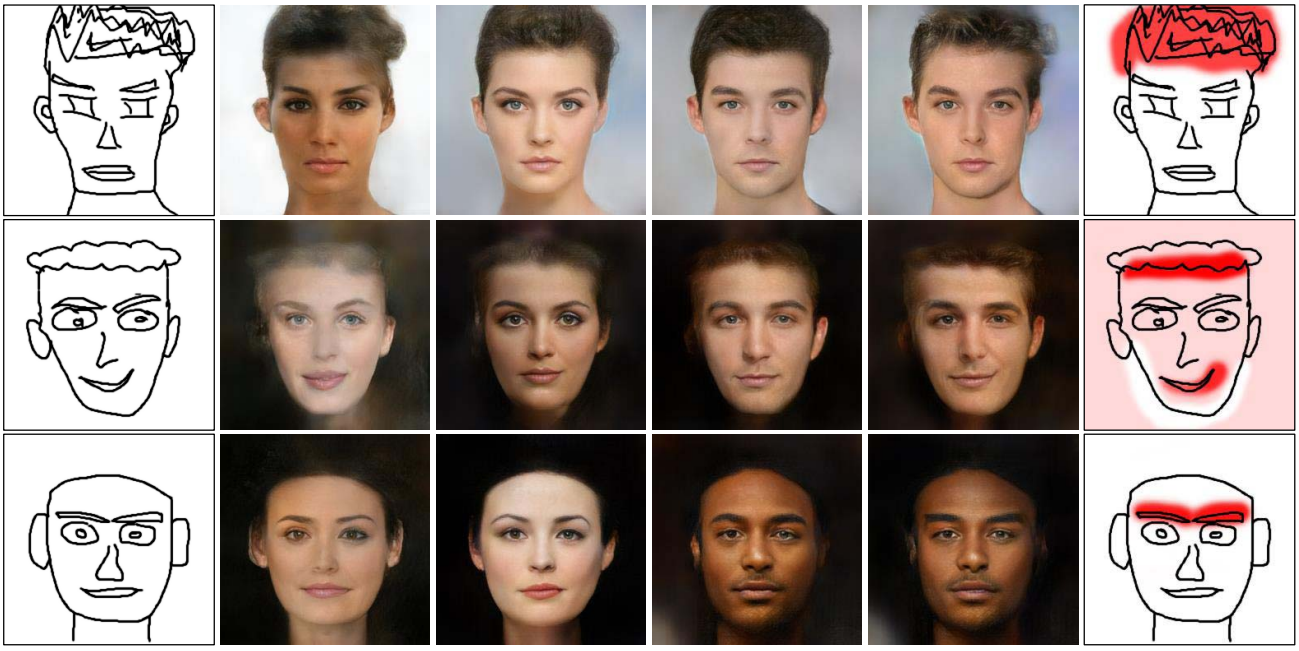


Fig. 17. Effect of facial attribute control and spatially non-uniform refinement level control. (a) User inputs. (b) Results by our base model [9]. (c) Results with class labels $c1$ where $c1$ is set to match the attributes in (b). (d) Results with class labels $c2$ where we manually select $c2$ that best matches the original sketches. (e) Results with class labels $c2$ and level map L to preserve local details in the input sketches. (f) Visualized level map L. The red (white) region in L uses a lower (higher) refinement level.

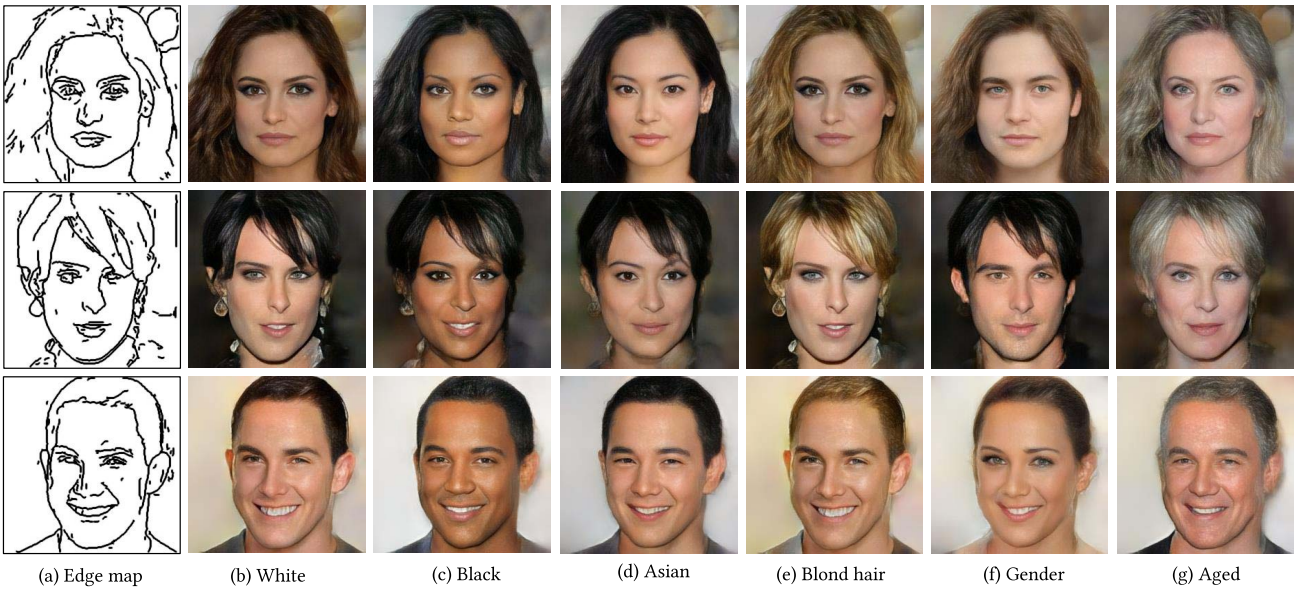


Fig. 18. Facial attribute control on edge maps from CelebA-HQ dataset.

TABLE IV
MEAN SQUARED ERROR OF THE RECONSTRUCTED EDGE MAPS

L	all-zero map	all-one map	map of zeros in the left and ones in the right	
			all-zero left part	all-one right part
MSE	0.005	0.490	0.025	0.489

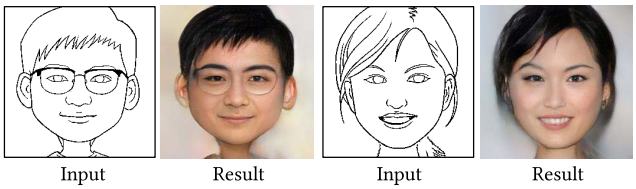


Fig. 19. Applications: cartoon-to-photo translation.

For simplicity, we use a binary label map where its left half part is all zeros, and the right half part is all ones. Then, we refine the edge maps of 1,000 testing images in CelebA-HQ [39] using our extended model with such label map. The mean squared errors (MSE) between the resulting S_{gen} and the original edge maps in the left part and the right

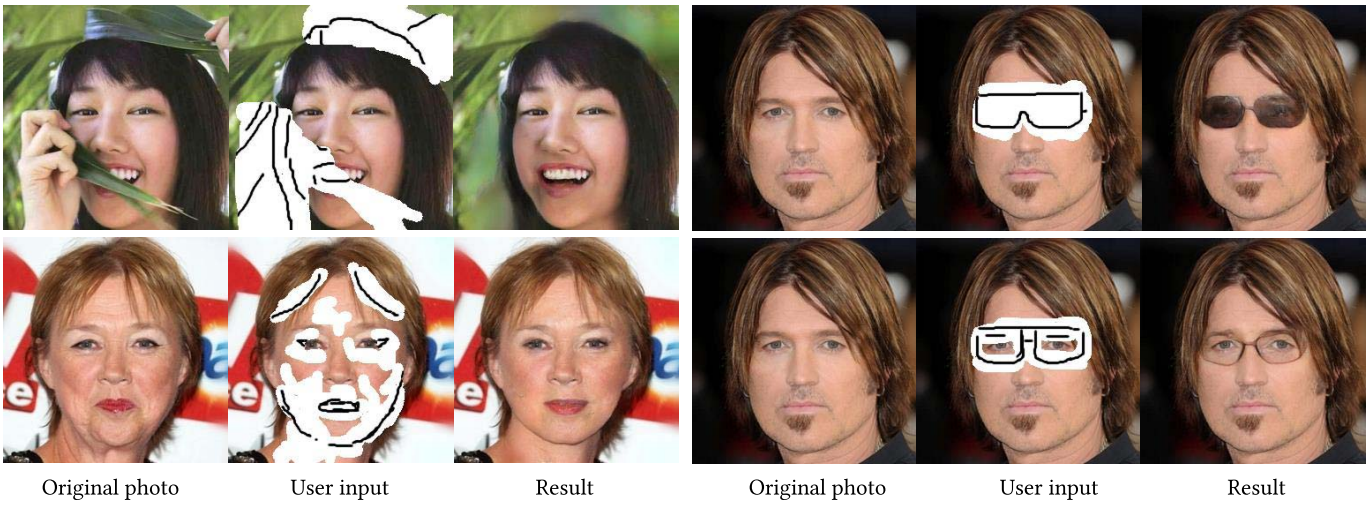


Fig. 20. Applications. From top to bottom, from left to right: object removal, rejuvenation and facial attribute editing.

part are reported in Table IV. The error is larger in the right part, meaning the model successfully adjusts the edges more drastically in this region. We further report the MSE using an all-one L and an all-zero L for reference, which is found to be matched with the MSE using the non-uniform L in the corresponding regions. It verifies that the refinement level of each position matches L's value in the corresponding position.

E. Applications

Fig. 18 presents the results of controlling facial attributes over edge maps of real-world face images, which realize facial attribute editings such as modifying the hair color or age.

In addition to the fine edge maps and rough sketches, our method can be applied to cartoons, which usually contains exaggerated facial structure such as big ears and thin necks. Fig. 19 presents two examples of cartoon-to-photo translation, where our method shows certain robustness.

Fig. 20 shows three applications of facial editing. First, our method can remove large occlusions using user guidance. Second, it allows users to perform "plastic surgery" digitally, including removing wrinkles and lifting the eye corners. Finally, our method can also edit fine-grained facial attributes such as adding glasses and further specifying their types.

Besides facial images, we further present our results on the handbag and shoe datasets [45], [46] and the Sketchy dataset [4] in Fig. 21. It can be seen that our method effectively designs novel handbags and shoes, and synthesizes realistic shoes from various real-world hand-drawn sketches.

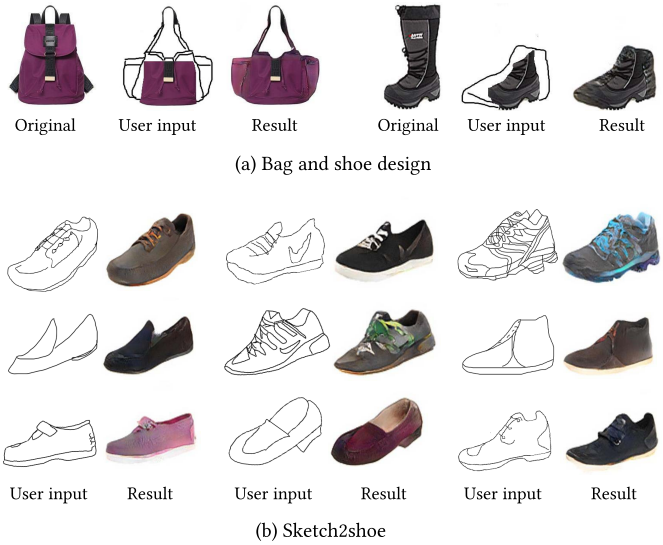


Fig. 21. Performance on other dataset. (a) Fashion design on handbag dataset [45] and shoe dataset [46]. (b) Shoe synthesis on Sketchy dataset [4].

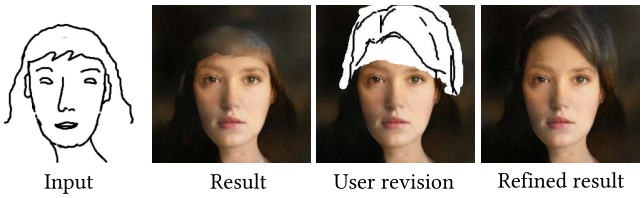


Fig. 22. User interaction for error revision.

F. Limitation and User Interaction

The robustness of our method to structural errors is limited by the maximum allowable radius R. Since large dilation will merge line details, make the coarse-to-fine mapping more ambiguous and difficult to learn, R cannot increase arbitrarily. Then, when the structural errors are large than R, our method cannot revise them. To address this problem, one possible solution is user interaction. Instead of synthesis in one step, users can always modify the input sketch and the temporary output until satisfied, as shown in Fig. 22.

VI. Conclusion

In this paper, we raise a new problem of controllable sketch-to-image translation, to adapt edge-based models to hand-drawn sketches, and present a novel dilation-based sketch

refinement framework. We show that the proposed drawable region modelling can effectively bridge the fine edges and rough sketches. An advantage is that it allows our method to work robustly on various sketches without the need for collecting real sketch data. We further demonstrate advanced user controllability in terms of refinement level, facial attribute and regional editing can be accomplished by the proposed comprehensive style-based feature modulations. Our method improves the generalizability and robustness of sketch-based image synthesis and editing. The network's ability to infer details from the dilated edges suggests a potential of building the relationship of two domains by degrading them into a shared rough domain, which might benefit more general research areas such as domain adaptation and transfer learning.

REFERENCES

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[2] T. Portenier, Q. Hu, A. Szabó, S. A. Bigdeli, P. Favaro, and M. Zwicker, "FaceShop: Deep sketch-based face image editing," *ACM Trans. Graph.*, vol. 37, no. 4, p. 99, 2018.

[3] Y. Jo and J. Park, "SC-FEGAN: Face editing generative adversarial network with user's sketch and color," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 1745–1753.

[4] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Trans. Graph.*, vol. 35, no. 4, p. 119, Jul. 2016.

[5] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C. C. Loy, "Sketch me that shoe," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 799–807.

[6] W. Chen and J. Hays, "SketchyGAN: Towards diverse and realistic sketch to image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9416–9425.

[7] R. Liu, Q. Yu, and S. Yu, "Unsupervised sketch to photo synthesis," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 36–52.

[8] Y. Lu, S. Wu, Y.-W. Tai, and C.-K. Tang, "Image generation from sketch constraint using contextual GAN," in *Proc. Eur. Conf. Comput. Vis.*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Berlin, Germany: Springer, 2018, pp. 205–220.

[9] S. Yang, Z. Wang, J. Liu, and Z. Guo, "Deep plastic surgery: Robust and controllable image editing with human-drawn sketches," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 601–617.

[10] 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.*, Jun. 2018, pp. 8798–8807.

[11] J.-Y. Zhu *et al.*, "Toward multimodal image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 465–476.

[12] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[13] S. Yang, J. Liu, W. Wang, and Z. Guo, "TET-GAN: Text effects transfer via stylization and destylization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 1238–1245.

[14] 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)*, Jun. 2019, pp. 2337–2346.

[15] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "SEAN: Image synthesis with semantic region-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5104–5113.

[16] 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)*, Oct. 2017, pp. 2242–2251.

[17] M. Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 700–708.

[18] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Berlin, Germany: Springer, 2018, pp. 172–189.

[19] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 768–783.

[20] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5400–5409.

[21] T. Dekel, C. Gan, D. Krishnan, C. Liu, and W. T. Freeman, "Sparse, smart contours to represent and edit images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3511–3520.

[22] A. Ghosh *et al.*, "Interactive sketch & fill: Multiclass sketch-to-image translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 1171–1180.

[23] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[24] J. Sun, L. Yuan, J. Jia, and H.-Y. Shum, "Image completion with structure propagation," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 861–868, 2005.

[25] Y. Wexler, E. Shechtman, and M. Irani, "Space-time completion of video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 463–476, Mar. 2007.

[26] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.

[27] J. Liu, S. Yang, Y. Fang, and Z. Guo, "Structure-guided image inpainting using homography transformation," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3252–3265, Dec. 2018.

[28] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[29] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[30] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4471–4480.

[31] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6721–6729.

[32] C. Zheng, T.-J. Cham, and J. Cai, "Pluralistic image completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1438–1447.

[33] E. Simo-Serra, S. Iizuka, and H. Ishikawa, "Real-time data-driven interactive rough sketch inking," *ACM Trans. Graph.*, vol. 37, no. 4, p. 98, 2018.

[34] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[35] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[36] J. Johnson, A. Alahi, and F. F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Berlin, Germany: Springer, 2016, pp. 694–711.

[37] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[38] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. IEEE Int. Conf. Mach. Learn.*, Jul. 2017, pp. 2642–2651.

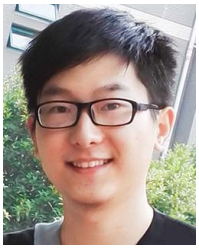
[39] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–12.

[40] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.

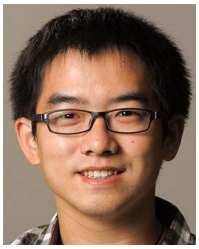
[41] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[42] A. Recasens, P. Kellnhofer, S. Stent, W. Matusik, and A. Torralba, "Learning to zoom: A saliency-based sampling layer for neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 51–66.

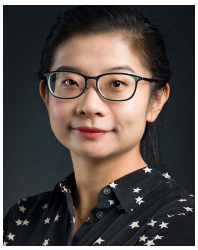
[43] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.
[44] S. Yang, Z. Wang, Z. Wang, N. Xu, J. Liu, and Z. Guo, "Controllable artistic text style transfer via shape-matching GAN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4442–4451.
[45] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 597–613.
[46] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 192–199.



Shuai Yang (Member, IEEE) received the B.S. and Ph.D. degrees (Hons.) in computer science from Peking University, Beijing, China, in 2015 and 2020, respectively. He was a Visiting Student with the National Institute of Informatics, Japan, from March 2017 to August 2017. He was a Visiting Scholar with Texas A&M University from September 2018 to September 2019. He is currently a Postdoctoral Research Fellow with the NTU AI Corporate Laboratory, Nanyang Technological University. His current research interests include image stylization and image inpainting. He received the IEEE ICME 2020 Best Paper Awards and the IEEE MMSP 2015 Top10% Paper Awards.



Zhangyang Wang (Senior Member, IEEE) received the B.E. degree from the University of Science and Technology of China (USTC) in 2012. From 2012 to 2016, he was a Ph.D. Student with the Electrical and Computer Engineering (ECE) Department, University of Illinois at Urbana-Champaign (UIUC), working with Prof. Thomas S. Huang. He was an Assistant Professor of CSE with Texas A&M University from 2017 to 2020. He is currently an Assistant Professor of ECE at UT Austin. He has coauthored over 100 papers, published two books, and one chapter. His research has been addressing machine learning, computer vision and optimization problems, and their interdisciplinary applications. He has received over 30 research awards. He is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT).



Jiaying Liu (Senior Member, IEEE) received the Ph.D. degree (Hons.) in computer science from Peking University, Beijing, China, in 2010.

She was a Visiting Scholar with the University of Southern California, Los Angeles, from 2007 to 2008. She was a Visiting Researcher with the Microsoft Research Asia in 2015 supported by the Star Track Young Faculties Award. She is currently an Associate Professor and a Peking University Boya Young Fellow with the Wangxuan Institute of Computer Technology, Peking University. She has authored over 100 technical articles in refereed journals and proceedings. She holds 50 granted patents. Her current research interests include multimedia signal processing, compression, and computer vision. She is a Senior Member of CSIG and CCF. She has served as a member for Multimedia Systems and Applications Technical Committee (MSA TC) and Visual Signal Processing and Communications Technical Committee (VSPC TC) in IEEE Circuits and Systems Society. She received the IEEE ICME-2020 Best Paper Awards and the IEEE MMSP-2015 Top10% Paper Awards. She has served as the Technical Program Chair for IEEE ICME-2021/ACM ICMR-2021, the Publicity Chair for IEEE ICME-2020/ICIP-2019, and the Area Chair for CVPR-2021/ECCV-2020/ICCV-2019. She has served as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and *JVCI* (Elsevier). She was the APSIPA Distinguished Lecturer from 2016 to 2017.

Zongming Guo (Member, IEEE) received the B.S. degree in mathematics and the M.S. and Ph.D. degrees in computer science from Peking University, Beijing, China, in 1987, 1990, and 1994, respectively.

He is currently a Professor with the Wangxuan Institute of Computer Technology, Peking University. His current research interests include video coding, processing, and communication.

Dr. Guo is an Executive Member of the China-Society of Motion Picture and Television Engineers. He was a recipient of the First Prize of the State Administration of Radio Film and Television Award in 2004, the First Prize of the Ministry of Education Science and Technology Progress Award in 2006, the Second Prize of the National Science and Technology Award in 2007, the Wang Xuan News Technology Award and the Chia Tai Teaching Award in 2008, the Government Allowance granted by the State Council in 2009, and the Distinguished Doctoral Dissertation Advisor Award of Peking University in 2012 and 2013.