Deep Plastic Surgery: Robust and Controllable Image Editing with Human-Drawn Sketches

Shuai Yang[1], Zhangyang Wang[2], Jiaying Liu[1(✉)], and Zongming Guo[1]

[1] Wangxuan Institute of Computer Technology, Peking University, Beijing, China
{williamyang,liujiaying,guozongming}@pku.edu.cn
[2] Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, USA
atlaswang@utexas.edu

Abstract. Sketch-based image editing aims to synthesize and modify photos based on the structural information provided by the human-drawn sketches. Since sketches are difficult to collect, previous methods mainly use edge maps instead of sketches to train models (referred to as edge-based models). However, human-drawn sketches display great structural discrepancy with edge maps, thus failing edge-based models. Moreover, sketches often demonstrate huge variety among different users, demanding even higher generalizability and robustness for the editing model to work. In this paper, we propose *Deep Plastic Surgery*, a novel, robust and controllable image editing framework that allows users to interactively edit images using hand-drawn sketch inputs. We present a sketch refinement strategy, as inspired by the coarse-to-fine drawing process of the artists, which we show can help our model well adapt to casual and varied sketches without the need for real sketch training data. Our model further provides a refinement level control parameter that enables users to flexibly define how "reliable" the input sketch should be considered for the final output, balancing between sketch faithfulness and output verisimilitude (as the two goals might contradict if the input sketch is drawn poorly). To achieve the multi-level refinement, we introduce a style-based module for level conditioning, which allows adaptive feature representations for different levels in a singe network. Extensive experimental results demonstrate the superiority of our approach in improving the visual quality and user controllablity of image editing over the state-of-the-art methods. Our project and code are available at https://github.com/TAMU-VITA/DeepPS.

Keywords: Image editing · Sketch-to-image translation · User control

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58555-6_36) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12360, pp. 601–617, 2020.
https://doi.org/10.1007/978-3-030-58555-6_36

1 Introduction

Human-drawn sketches reflect people's abstract expression of objects. They are highly concise yet expressive: usually several lines can reflect the important morphological features of an object, and even imply more semantic-level information. Meanwhile, sketches are easily editable: such an advantage is further amplified by the increasing popularity of touch-screen devices. Sketching thus becomes one of the most important ways that people illustrate their ideas and interact with devices. Motivated by the above, a series of sketch-based image synthesis and editing methods have been proposed in recent years. The common main idea underlying these methods is to train an image-to-image translation network to map a sketch to its corresponding color image. That can be extended to an image completion task where an additional mask is provided to specify the area for modification. These methods enable novice users to edit the photo by simply drawing lines, rather than resorting complicated tools to process the photo itself.

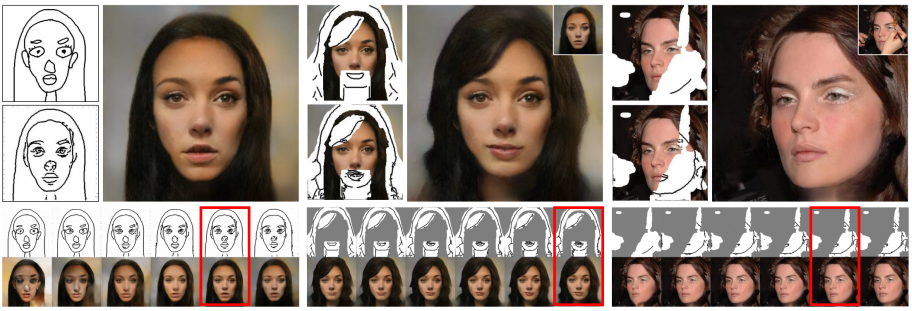


Fig. 1. Our Deep Plastic Surgery framework allows users to synthesize (left) and edit (middle, right) photos based on hand-drawn sketches. Our model is robust to tolerate the drawing errors and achieves the controllability on sketch faithfulness. For each group, we show the user input and our refined sketch in the left column, and the final output in the right column with the original photo in the upper right corner. The bottom row shows our results under an increasing refinement level, with a red box to indicate the user selection. Note that our model requires no real sketch for training. (Color figure online)

Due to the difficulty of collecting pairs of sketches and color images as training data, existing works [11,12,22] typically exploit edge maps (detected from color images) as "surrogates" for real sketches, and train their models on the paired edge-photo datasets. Despite certain success in shoe, handbag and face synthesis, edge maps look apparently different from the human drawings, the latter often being more causal, varied or even wild. As a result, those methods often generalize poorly when their inputs become human-drawn sketches, limiting their real-world usage. To resolve this bottleneck, researchers have studied edge pre-processing [22], yet with limited performance improvement gained so

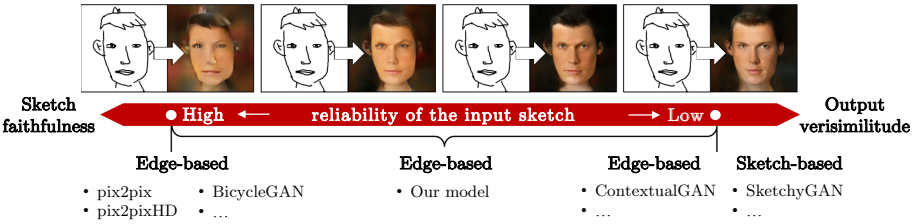


Fig. 2. Illustration of the sketch-to-image translation spectrum. Our model differs from existing models in that we allow users to define how "reliable" the input sketch should be considered for the final output, thus balancing between sketch faithfulness and output verisimilitude, which has not been well studied in previous approaches. As edge-based models, ContexualGAN [20] and our model realize verisimilitude without real sketch data for training, and our model further achieves controllability and efficiency.

(a) Edge map (ℓ=0) (b) Fine sketch (ℓ=0.25) (c) Rough sketch (ℓ=0.75) (d) Poor sketch (ℓ=1)

Fig. 3. Our model works robustly on various sketches by setting refinement level ℓ adaptive to the quality of the input sketches, *i.e.*, higher ℓ for poorer sketches.

far. Some human-drawn sketch datasets have also been collected [24,39] to train sketch-based models [3,18]. However, the collection is too laborious to extend to larger scales or to meet all data subject needs.

As a compromise, it is valuable to study the adaption of edge-based models to the sketches. ContextualGAN [20] presents an intuitive solution. It retrieves the nearest neighbor of the input sketch from the learned generative edge-image manifolds, which relaxes the sketch constraint to trade for the image naturalness. However, neither edge-based models [11,12,22] nor ContextualGAN [20] allows for any user controllability on the *sketch faithfulness*, *i.e.*, to what extent we should stick to the given sketch? The former categories of methods completely hinge on the input sketch even it might yield highly unnatural outputs; while the latter mainly searches from natural manifolds and may produce visually disparate results from the sketch specification. That leaves little room for users to calibrate between freedom of sketching and the overall image verisimilitude: an important desirable feature for interactive photo editing.

In view of the above, we are motivated to investigate a new problem of controllable sketch-based image editing, that can work robustly on varied human-drawn sketches. Our main idea is to refine the sketches to approach the structural features of the edge maps, therefore avoiding the tedious collection of sketch data for training, while enabling users to control the refinement level freely. Figure 1 intuitively demonstrates our task: to improve the model's robustness to various sketch inputs by allowing users to navigate around editing results under different

refinement levels and select the most desired one. The challenge of this problem lies in two aspects. First, in the absence of real sketches as reference, we have no paired or unpaired data to directly establish a mapping between sketches and edge maps. Second, in order to achieve controllability, it is necessary to extend the above mapping to a multi-level progress, which remains to be an open question. Please refer to Fig. 2 to get a sense of the difference between the proposed controllable model from common sketch-to-image translation models.

In this paper, we present *Deep Plastic Surgery*, a novel sketch-based image editing framework to achieve both **robustness** on hand-drawn sketch inputs, and the **controllability** on sketch faithfulness. Our key idea arises from our observation on the coarse-to-fine drawing process of the human artists: they first draw coarse outlines to specify the approximate areas where the final fine-level lines are located. Those are then gradually refined to converge to the final sharper lines. Inspired by so, we propose a dilation-based sketch refinement method. Instead of directly feeding the network with the sketch itself, we only specify the approximate region covering the final lines, created by edge dilation, which forces the network to find the mapping between the coarse-level sketches and fine-level edges. The level of coarseness can be specified and adjusted by setting the dilation radius. Finally, we treat sketches under different coarse levels as different stylized versions of the fine-level lines, and use the scale-aware style transfer to recover fine lines by removing their dilation-based styles. Our method only requires color images and their edge maps to train and can adapt to diversified sketch input. It can work as a plug-in for existing edge-based models, providing refinement for their inputs to boost their performance. Figure 3 shows an overall performance of our method on various sketches.

Our contributions are summarized as three-folds:

- We explore a new problem of controllable sketch-based image editing, to adapt edge-based models to human-drawn sketches, where the users have the freedom to balance the sketch faithfulness with the output verisimilitude.
- We propose a sketch refinement method using coarse-to-fine dilations, following the drawing process of artists in real world.
- We propose a style-based network architecture, which successfully learns to refine the input sketches into diverse and continuous levels.

2 Related Work

Sketch-Based Image Synthesis. Using the easily accessible edge maps to simulate sketches, edge-based models [6,11,25] are trained to map edges to their corresponding photos. By introducing masks, they are extended to image inpainting tasks to modify the specified photo areas [12,22,38] or provide users with sketch recommendations [7]. However, the drastic structural discrepancy between edges and human-drawn sketches makes these models less generalizable to sketches. As sketches draw increasing attentions and some datasets [24,39] are released, the discrepancy can be narrowed [3,18]. But existing datasets are far from enough and collecting sketches in large scale is still too expensive.

The most related method to our problem setting is ContextualGAN [20] that also aims to adapt edge-based models to sketches. It solves this problem by learning a generative edge-image manifold through GANs, and searching nearest neighbors to the input sketch in this manifold. As previously discussed, ContextualGAN offers no controllablity, and the influence of the sketch input might be limited for the final output. As can be seen in Fig. 10, ContextualGAN cannot well preseve some key sketch features. Besides, the nearest neighbor search costs time-consuming iterative back-propagation. It also relies on the generative manifolds provided by GANs, which can become hard to train as image resolution grows higher. Thus, results reported in [20] are of a limited 64×64 size. By comparison, our method is able to refine 256×256 sketches in a fast feed-forward way, with their refinement level controllable to better preserve the shape and details of the sketches and to facilitate flexible and user-friendly image editing.

Image-to-Image Translation. Image-to-image translation networks have been proposed to translate an image from a source domain into a target domain. Isola *et al.* [11] designed a general image-to-image translation framework named pix2pix to map semantic label maps or edge maps into photos. Follow-ups involve the diversification of the generated images [41], high-resolution translation [30], and multi-domain translation [4,34,35]. This framework requires that images in two domains exist as pairs for training. Zhu *et al.* [40] suggested a cycle consistency constraint to map the translated image back to its original version, which successfully trained CycleGAN on unpaired data. By assuming a shared latent space across two domains, UNIT [17] and MUNIT [10] are proposed upon CycleGAN to improve the translation quality and diversity.

Image Inpanting. Image inpanting aims to reconstruct the missing parts of an image. Early work [2] smoothly propagates pixel values from the known region to the missing region. To deal with large missing areas, examplar-based methods are proposed to synthesize textures by sampling pixels or patches from the known region in a greedy [5,28] or global [1,16,31] manner. However, the aforementioned methods only reuse information of known areas, but cannot create unseen content. In parallel, data-driven methods [8,26,29] are proposed to achieve creative image inpanting or extrapolation by retrieving, aligning and blending images of similar scenes from external data. Recent models such as Context Encoder [21] and DeepFill [37,38] build upon the powerful deep neural networks to leverage the extra data for semantic completion, which supports fast intelligent image editing for high-resolution images [33] and free-form masks [38].

3 The Deep Plastic Surgery Algorithm

As illustrated in Fig. 4, given an edge-based image editing model F trained on edge-image pairs $\{S_{gt}, I_{gt}\}$, our goal is to adapt F to human-draw sketches through a novel sketch refinement network G that can be trained without sketch data. G aims to refine the input sketch to match the fine edge maps S_{gt}. The output is then fed into F to obtain final editing results. Our model is further

conditioned by a control parameter $\ell \in [0, 1]$ indicating the refinement level, where larger ℓ corresponds to greater refinement.

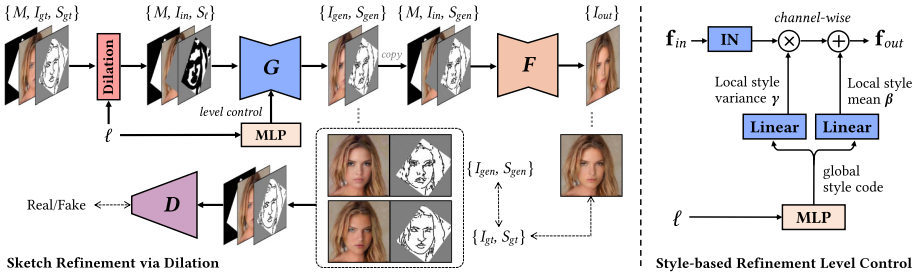


Fig. 4. Framework overview. A novel sketch refinement network G is proposed to refine the rough sketch S_ℓ modelled as dilated drawable regions to match the fine edges S_{gt}. The refined output S_{gen} is fed into a pretrained edge-based model F to obtain the final editing result I_{out}. A parameter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

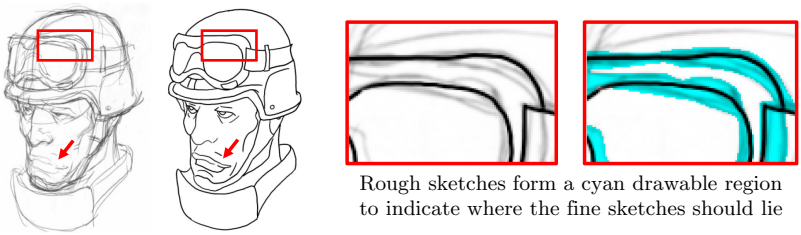


Fig. 5. Rough sketch (left) to fine sketch (middle). The sketches in the red boxes are enlarged and overlayed on the right. Image is copyrighted by Krenz Cushart [27]. (Color figure online)

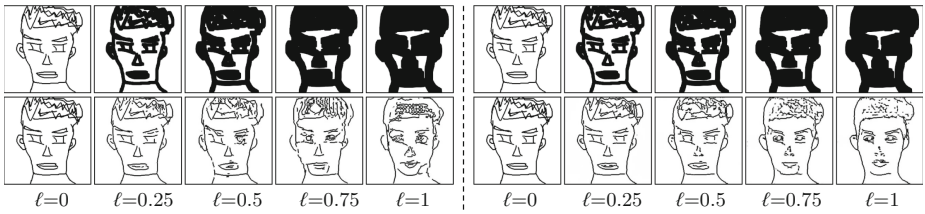
3.1 Sketch Refinement via Dilation

Our sketch refinement method is inspired by the coarse-to-fine drawing process of human artists. As shown in Fig. 5, artists usually begin new illustrations with inaccurate rough sketches with many redundant lines to determine the shape of an object. These sketches are gradually fintuned by merge lines, tweaking details and fixing mistakes to obtain the final line drawings. When overlaying the final lines on the rough sketches, we find that the redundant lines in the rough sketches form a drawable region to indicate where the final lines should lie (tinted in cyan in Fig. 5). Thus *the coarse-to-fine drawing process is essentially a process of continuously reducing the drawable region.*

Based on the observation, we define our sketch refinement as an image-to-image translation task between rough and fine sketches, where in our problem,

(a) Input edge (b) Deform + discard (c) Dilate (d) Difference map

Fig. 6. Rough sketch synthesis. (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.



(a) pix2pix (b) Our multi-level model

Fig. 7. Sketch refinement at different level ℓ. Top row: S_ℓ with different dilation radii. Bottom row: (a) Refinement results by pix2pix [11] trained separately for each level. (b) Refinement results by our proposed single model with multi-level control.

fine sketches S_{gt} are edge maps extracted from I_{gt} using HED edge detector [32] and *rough sketches are modelled as drawable regions* $\Omega(S_{gt})$ completely covering S_{gt}. In the following, we present our dilation-based drawable region generation method to automatically generate $\Omega(S_{gt})$ based on S_{gt} to form our training data.

Rough Sketch Data Generation. The pipeline of our drawable region generation is shown in Fig. 6. The main idea is to expand lines into areas by dilation operations used in mathematical morphology. However, directly learning to translate a dilated line back to itself will only make the network to simply extract the skeleton centered at the region without refining the sketches. Thus the fine lines are first randomly deformed before dilation. Supposing the radius of dilation is r, then we limit the offset of each pixel after deformation to no more than r, so that the ground truth fine lines are not centered at the drawable region but still fully covered by it, as shown in Fig. 6(d). In addition, noticing that artists will also infer new structures or details from the draft (see the upper lip pointed by the red arrow of Fig. 5), we further discard partial lines by removing random patches from the full sketches. By doing so, our network is motivated to learn to complete the incomplete structures such as the cyan lines in Fig. 6(d). Line deformation and discarding are only applied during the training phase.

Leveraging our dilation-based drawable region generation algorithm, sufficient paired data $\{\Omega(S_{gt}), S_{gt}\}$ is obtained. Intuitively, larger drawable regions

provide more room for line-fintuning, which means a higher refinement level. To verify our idea of coarse-to-fine refinement, we train a basic image-to-image translation model of pix2pix [11] to map $\Omega(S_{gt})$ to S_{gt} and use a separate model for each dilation radius. As shown in Fig. 7(a), the rough facial structures are refined and a growing refinement is observed as the radius increases. This property makes it possible for convenient sketch editing control. In the next section, we will detail how we incorporate sketch refinement into one single model with effective level control, whose overall performance is illustrated in Fig. 7(b). The advantage is that coarse-level refinement can benefit from the learned robust fine-level features, thus achieving better performance.

3.2 Controllable Sketch-Based Image Editing

In our image editing task, we have a target photo I_{gt} as input, upon which a mask M is given to indicate the editing region. Users draw sketches S to serve as a shape guidance for the model to fill the masked region. The model will adjust S so that it better fits the contextual structures of I_{gt}, with the refinement level determined by a parameter ℓ.

Our training requires no human-drawn sketches. Instead, we use edge maps $S_{gt} = \text{HED}(I_{gt})$ [32] and generate their corresponding drawable regions $\Omega(S_{gt})$. As analyzed in Sect. 3.1, the refinement level is positively correlated with the dilation radius r. Therefore, we incorporate ℓ in the drawable region generation process (denoted as $\Omega_\ell(\cdot)$) to control r, where $r = \ell R$ with R the maximum allowable radius. The final drawable region with respect to ℓ takes the form of $S_\ell = \Omega_\ell(S_{gt}) \odot M$ where $\odot$ is the element-wise multiplication operator. Then we are going to train G to map S_ℓ back to the fine S_{gt} based on the contextual condition $I_{in} = I_{gt} \odot (\mathbf{1} - M)$, the spatial condition M and the level condition ℓ. Figure 4 shows an overview of our network architecture. G receives a concatenation of I_{in}, S_ℓ and M, with middle layers controlled by ℓ, and yields a four-channel tensor: the completed RGB channel image I_{gen} and the refined one channel sketch S_{gen}, *i.e.*, $(I_{gen}, S_{gen}) = G(I_{in}, S_\ell, M, \ell)$. Here, we task the network with photo generation to enforce the perceptual guidance on the edge generation. It also enables our model to work independently if F is unavailable. Finally, a discriminator D is added to improve the results through adversarial learning.

Style-Based Refinement Level Control. As we will show later, conditioning by label concatenation or feature interpolation [36] fails to properly condition G about the refinement level. Inspired by AdaIN-based style transfer [9] and image generation [15], we propose an effective style-based control module to address this issue. Specifically, sketches at different coarse levels can be considered to have different styles. And G is tasked to destylize them to obtain the original S_{gt}. In AdaIN [9], styles are modelled as the mean and variance of the features and are transferred via distribution scaling and shifting (*i.e.*, normalization+denormalization). Note that the same operation can also be used for its reverse process, *i.e.*, destylization. To this end, as illustrated by Fig. 4, we

propose to use a multi-layer perceptron to decode the condition ℓ into a global style code. For each convolution layer expect the first and the last ones in G, we have two affiliated linear layers to map the style code to the local style mean and variance for AdaIN-based destylization.

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

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{I_{gt},M,\ell}[\|I_{gen} - I_{gt}\|_1 + \|S_{gen} - S_{gt}\|_1 + \|I_{out} - I_{gt}\|_1], \tag{1}$$

where $I_{out} = F(I_{in}, S_{gen}, M)$ is the ultimate output in our problem. Here the quality of I_{out} is also considered to adapt G to the pretrained F in an end-to-end manner. Besides, perceptual loss $\mathcal{L}_{\text{perc}}$ [13] to measure the semantical similarity of the photos is computed as

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

where $\Phi_i(x)$ is the feature map of x in the i-th layer of VGG19 [23] and λ_i is the layer weight. Finally, we use hinge loss as our adversarial objective function:

$$\mathcal{L}_G = -\mathbb{E}_{I_{gt},M,\ell}[D(I_{gen}, S_{gen}, M)], \tag{3}$$

$$\mathcal{L}_D = \mathbb{E}_{I_{gt},M,\ell}[\sigma(\tau + D(I_{gen}, S_{gen}, M))] + \mathbb{E}_{I_{gt},M}[\sigma(\tau - D(I_{gt}, S_{gt}, M))], \tag{4}$$

where τ is a margin parameter and σ is ReLU activation function.

Realistic Sketch-to-Image Translation. Under the extreme condition of $M = \mathbf{1}$, I_{gt} is fully masked out and our problem becomes a more challenging sketch-to-image translation problem. We experimentally find that the result will degrade without any contextual cues from I_{gt}. To solve this problem, we adapt our model by removing the I_{in} and M inputs, and train a separate model specifically for this task, which brings obvious quality improvement.

4 Experimental Results

4.1 Implementation Details

Dataset. We use CelebA-HQ dataset [14] with edge maps extracted by HED edge detector [32] to train our model. The masks are generated as the randomly rotated rectangular regions following [22]. To make a fair comparison with ContextualGAN [20], we also train our model on CelebA dataset [19].

Network Architecture. Our generator G utilizes the Encoder-ResBlocks-Decoder [13] with skip connections [11] to preserve the low-level information. Each convolutional layer is followed by AdaIN layer [9] except the first and the last layer. The discriminator D follows the SN-PatchGAN [38] for stable and fast training. Finally, we use pix2pix [11] as our edge-based baseline model F.

Network Training. We first train our network with $\ell = 1$ for 30 epoches, and then train with uniformly sampled $\ell \in [0, 1]$ for 200 epoches. The maximum

Fig. 8. Comparison with state-of-the-art methods on face edting. (a) Input photos, masks and sketches. (b) DeepFillv2 [38]. (c) SC-FEGAN [12]. (d) Our results with $\ell = 0$. (e) Our results with $\ell = 1$. (f) SC-FEGAN using our refined sketches as input.

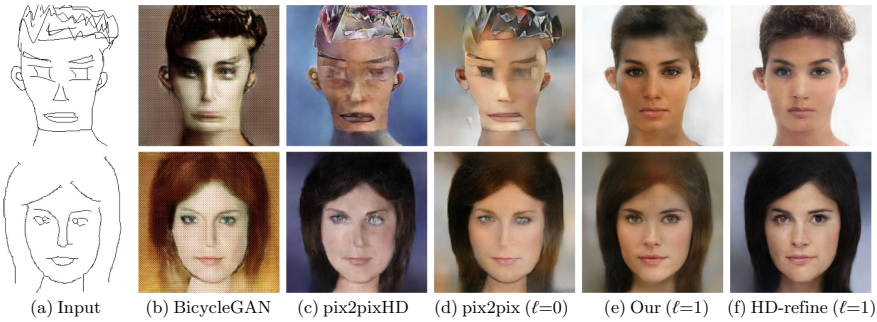


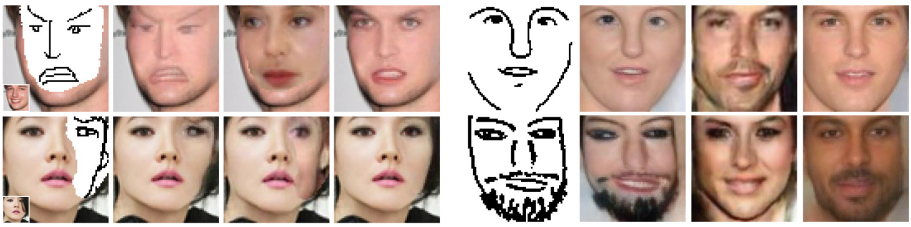
Fig. 9. Comparison with state-of-the-art methods on face synthesis. (a) Input human-drawn sketches. (b) BicycleGAN [41]. (c) pix2pixHD [30]. (d) pix2pix [11]. (e) Our results with $\ell = 1$. (f) pix2pixHD using our refined sketches as input.

allowable dilation radius is set to $R = 10$ for CelebA-HQ dataset [14] and $R = 4$ for CelebA dataset [19]. For all experiments, the weight for $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{perc}}$, $\mathcal{L}_G$ and $\mathcal{L}_{\text{D}}$ are 100, 1, 1 and 1, respectively. To calculate $\mathcal{L}_{\text{perc}}$, we use the conv2_1 and conv3_1 layers of the VGG19 [23] weighted by 1 and 0.5, respectively. For hinge loss, we set τ to 10 and 1 for G and F, respectively.

Please refer to our supplementary material and project page for more details.

4.2 Comparisons with State-of-the-Art Methods

Face Editing and Synthesis. Figure 8 presents the qualitative comparison on face editing with two state-of-the-art inpainting models: DeepFillv2 [38] and SC-FEGAN [12]. 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 inpainting models, DeepFillv2 fails to repair the fine-scale facial structures well, indicating the necessity of user guidance. SC-FEGAN, on the



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

Fig. 10. Comparison with ContextualGAN [20] on face editing and face synthesis.

Table 1. User preference ratio of state-of-the-art methods.

Task	Face Editing			Face Synthesis			Face Synthesis*	
Method	DeepFillv2	SC-FEGAN	Ours	BicycleGAN	pix2pixHD	Ours	ContextualGAN	Ours
Score	0.032	0.238	**0.730**	0.024	0.031	**0.945**	0.094	**0.906**

* ContextualGAN is designed for image synthesis on 64×64 images. We have tried to extend ContextualGAN to 256×256. However, due to the inherent difficulty of training noise-to-image GAN on high resolution, ContextualGAN easily falls into model collapse with poor results. Therefore, we make a separate comparison with it on 64×64 images in the user study.

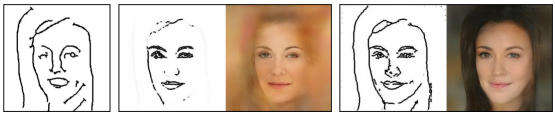
other hand, totally follows the inaccurate sketch and yields weird faces. Similar results can be found in the output of F when $\ell = 0$. By using a large refinement level ($\ell = 1$), the facial details become more natural and realistic. Finally, as an ablation study to indicate the importance of sketch-edge input adaption, we directly feed SC-FEGAN with our refined sketch (without fine-tuning upon SC-FEGAN), and observe improved results of SC-FEGAN.

Figure 9 shows the qualitative comparison on face synthesis with two state-of-the-art image-to-image translation models: BicycleGAN [41] and pix2pixHD [30]. As expected, both models as well as F (pix2pix [11]) synthesize facial structures that strictly match the inaccurate sketch inputs, producing poor results. Our model takes sketches as "useful yet flexible" constraints, and strikes a good balance between authenticity and consistency with the user guidance.

Comparison with ContextualGAN. As the most related work that accepts weak sketch constraint as our model, we further compare with it in this section. For face editing task, we implement ContextualGAN and adapt it to the completion task by additionally computing the appearance similarity between the known part of the photo during the nearest neighbor search. As shown in Fig. 10(a), the main downside of ContextualGAN is the distinct inpainting boundaries, likely due to that the learned generative manifold does not fully depict the real facial distribution. By comparison, our method produces more natural results. Figure 10(b) shows the sketch-to-image translation results, where the results of ContextualGAN are directly imported from the original paper. As can be seen, although realistic, the results of ContextualGAN lose certain

(a) No refinement (b) Baseline (c) Baseline+DF+DC (d) Full model

Fig. 11. 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 and discarding. (d) Edge dilation with multiple radii + line deformation and discarding.



(a) Input (b) w/o adaptation (c) w/ adaptation

Fig. 12. Effect of adaptation to F.

attributes associated with the input such as the beard. It might be because the learned generative manifolds collapse for some uncommon attributes. As another possible cause, the nearest neighbor search might sometimes travel too far over the manifold, and results in found solutions less relevant to the initial points provided by user sketches. Our method preserves these attributes much better.

In terms of efficiency, for 64×64 images in Fig. 10, our implemented ContextualGAN requires about 7.89 s per image with a GeForce GTX 1080 Ti GPU, while the proposed feed-forward method only takes about **12 ms per image**.

Quantitative Evaluation. To better understand the performance of the compared methods, we perform user studies for quantitative evaluations. A total of 28 face editing and 38 face synthesis cases are used and participants are asked to select which result best balances the sketch faithfulness with the output verisimilitude. We finally collect totally 1,320 votes from 20 subjects and demonstrate the preference scores in Table 1. The study shows that our method receives most votes for both sketch detail preservation and output naturalness.

4.3 Ablation Study

In this section, we perform ablation studies to verify our model design. We test on the challenging sketch-to-image translation task for better comparison.

Rough Sketch Modelling. We first examine the effect of our dilation-based sketch modelling, which is the key of our sketch refinement. In Fig. 11, we perform a comparison between different rough sketch models. The dilation prompts the network to infer the facial details. Then the line deformation and discarding force the network to further infer and complete the accurate facial structures. In Fig. 11(d), we observe an improvement brought by learning multiple refinement

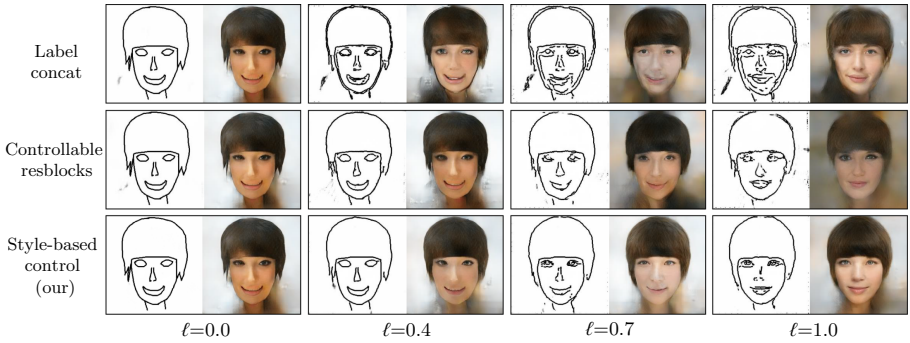


Fig. 13. Visual comparison on label conditioning.

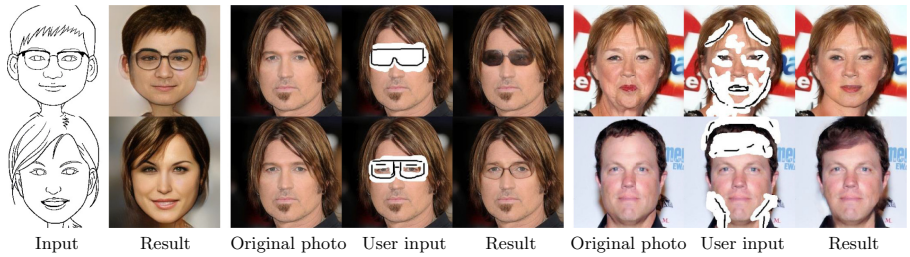


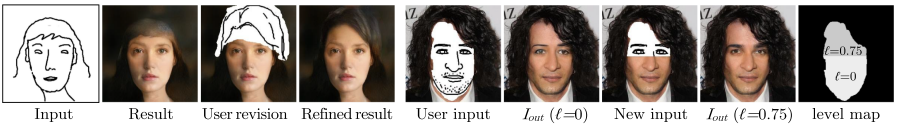
Fig. 14. Applications. More applications can be found in the supplemental material.

levels in one model over single level per model. The reason might be that coarse-level refinement can benefit from the learned more robust fine-level features.

Adaptation to F. Our generator G is trained together with a fixed F, which adapts G to F to improve the quality of the ultimate output. To verify the effect of the adaptation, we train a model without the loss terms related to I_{out} in Eqs. (1) and (2). Figure 12 presents the comparison of our model with and without adaptation. The sketch result without adaptation has its structure refined but some lines become indistinct. The reason might be the low proportion of the line region in the sketch. Through adaptation, G is motivated to generate sketches that are fully perceivable by F, which actually acts as a sketch-version perceptual loss [13], resulting in distinct lines and high-quality photos.

Refinement Level Control. We compare the proposed style-based conditioning with label concatenation and controllable resblock [36] in Fig. 13. Label concatenation yields stacking lines like those in draft sketches. Controllable resblock generates cleaner lines but still rough facial details. Our style-based conditioning surpasses controllable resblock in adaptive channel-wise control, which provides strongest results in both well-structured sketches and realistic photos.

Fig. 15. Applications on handbag and shoe design.



(a) User interactive revision (b) Spatially non-uniform refinement

Fig. 16. User interaction for error revision and spatially non-uniform refinement.

4.4 Applications

Figure 14 shows various results with facial sketch inputs. Our model shows certain robustness on realistic photo rendering from cartoons. Our model can also edit facial attributes such as adding glasses. Finally, users can purposely perform "plastic surgery" digitally, such as removing wrinkles, lifting the eye corners. Alternatively, amateurs can intuitively edit the face with fairly coarse sketches to provide a general idea, such as face-lifting and bangs, and our model will tolerate the drawing errors and suggest a suitable "surgery" plan.

We further present our results on the handbag and shoe datasets [11] and Sketchy dataset [24]. The results are shown in Fig. 15, where our model can effectively design handbags and shoes.

4.5 Limitation and User Interaction

User Interactive Revision. While our approach has generated appealing results, limitations still exist. Our method cannot revise the structural error that exceeds the maximum allowable radius. This problem can be possibly solved by user interaction, where users can modify the input sketch when the output is still unsatisfactory under the maximum refinement level as shown in Fig. 16(a).

Spatially Non-uniform Refinement. Another limitation is that, when ℓ is large, the dilation operation will merge lines that are close to each other, which inevitably loses some structural details. One solution is to use adaptive spatially varied dilation radii. In addition, the accuracy of the structure can vary within one sketch, which also demands spatially non-uniform sketch refinement for more flexible controllability. Our model can be easily extended to spatially non-uniform refinement with user interaction. As shown in Fig. 16(b), the user first uses a low ℓ on the whole mask region to better comply with the structure guidance of the nose, mouth and stubble. Then, user can edit the mask and further improve the verisimilitude of the eye region with a high ℓ. It allows users to improve the overall facial structure while achieving better detail preservation.

5 Conclusion

In this paper, we raise a new a new problem of controllable sketch-based image editing, to adapt edge-based models to human-drawn sketches, and present a novel dilation-based sketch refinement method. Modelling the rough sketch as a drawable region via edge dilation, the network is effectively trained to infer accurate structural information. Leveraging the idea of style transfer, our network is able to undo the edge dilation of different levels in a destylization manner for multi-level refinement control. We validate by experiments the effectiveness and robustness of our method. Serving as a plug-in, our model can greatly improve the performance of edge-based models on the sketch inputs.

Acknowledgement. This work was supported in part by National Natural Science Foundation of China under contract No. 61772043, and in part by Beijing Natural Science Foundation under contract No. L182002 and No. 4192025. The research of Z. Wang was partially supported by NSF Award RI-1755701. This work was supported by China Scholarship Council.

References

1. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009)
2. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of ACM SIGGPRAH, pp. 417–424 (2000)
3. Chen, W., Hays, J.: SketchyGAN: towards diverse and realistic sketch to image synthesis. In: Proceedings of IEEE International Conference Computer Vision and Pattern Recognition, pp. 9416–9425 (2018)
4. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of IEEE International Conference Computer Vision and Pattern Recognition (2018)
5. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. **13**(9), 1200–1212 (2004)
6. Dekel, T., Gan, C., Krishnan, D., Liu, C., Freeman, W.T.: Sparse, smart contours to represent and edit images. In: Proc. IEEE International Conference Computer Vision and Pattern Recognition, pp. 3511–3520 (2018)
7. Ghosh, A., et al.: Interactive sketch & fill: multiclass sketch-to-image translation. In: Proceedings of International Conference Computer Vision (2019)
8. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Trans. Graph. **26**(3), 4 (2007)
9. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: roceedings of International Conference Computer Vision, pp. 1510–1519 (2017)
10. Huang, X., Liu, M.-Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11207, pp. 179–196. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01219-9_11

11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of IEEE International Conference Computer Vision and Pattern Recognition, pp. 5967–5976 (2017)
12. Jo, Y., Park, J.: SC-FEGAN: face editing generative adversarial network with user's sketch and color. In: Proceedings of International Conference Computer Vision (2019)
13. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
14. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: Proceedings of International Conference, Learning Representations (2018)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of IEEE International Conference Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
16. Liu, J., Yang, S., Fang, Y., Guo, Z.: Structure-guided image inpainting using homography transformation. IEEE Trans. Multimedia **20**(12), 3252–3265 (2018)
17. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems, pp. 700–708 (2017)
18. Liu, R., Yu, Q., Yu, S.: An unpaired sketch-to-photo translation model (2019). arXiv:1909.08313
19. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738 (2015)
20. Lu, Y., Wu, S., Tai, Y.-W., Tang, C.-K.: Image generation from sketch constraint using contextual GAN. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11220, pp. 213–228. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01270-0_13
21. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings IEEE International Conference Computer Vision and Pattern Recognition, pp. 2536–2544 (2016)
22. Portenier, T., Hu, Q., Szabo, A., Bigdeli, S.A., Favaro, P., Zwicker, M.: Faceshop: deep sketch-based face image editing. ACM Trans. Graph. **37**(4), 99 (2018)
23. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
24. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: learning to retrieve badly drawn bunnies. ACM Trans. Graph. **35**(4), 119:1–119:12 (2016)
25. Sangkloy, P., Lu, J., Fang, C., Yu, F., Hays, J.: Scribbler: controlling deep image synthesis with sketch and color. In: Proceedings IEEE International Conference Computer Vision and Pattern Recognition, pp. 5400–5409 (2017)
26. Shan, Q., Curless, B., Furukawa, Y., Hernandez, C., Seitz, S.M.: Photo uncrop. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 16–31. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_2
27. Simo-Serra, E., Iizuka, S., Ishikawa, H.: Real-time data-driven interactive rough sketch inking. ACM Trans. Graph. **37**(4), 98 (2018)
28. Sun, J., Yuan, L., Jia, J., Shum, H.Y.: Image completion with structure propagation. ACM Trans. Graph. **24**(3), 861–868 (2005)
29. Wang, M., Lai, Y., Liang, Y., Martin, R.R., Hu, S.M.: Biggerpicture: data-driven image extrapolation using graph matching. ACM Trans. Graph. **33**(6) (2014)

30. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proceedings of IEEE International Conference Computer Vision and Pattern Recognition (2018)
31. Wexler, Y., Shechtman, E., Irani, M.: Space-time completion of video. IEEE Trans. Pattern Anal. Mach. Intell. **3**, 463–476 (2007)
32. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of IEEE International Conference Computer Vision and Pattern Recognition, pp. 1395–1403 (2015)
33. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: Proceedings of IEEE International Conference Computer Vision and Pattern Recognition (2017)
34. Yang, S., Liu, J., Wang, W., Guo, Z.: TET-GAN: text effects transfer via stylization and destylization. Proc. AAAI Conf. Artif. Intell. **33**, 1238–1245 (2019)
35. Yang, S., Wang, W., Liu, J.: TE141K: artistic text benchmark for text effect transfer. IEEE Trans. Pattern Anal. Mach. Intell. **PP**(99), 1–15 (2020). https://doi.org/10.1109/TPAMI.2020.2983697
36. Yang, S., Wang, Z., Wang, Z., Xu, N., Liu, J., Guo, Z.: Controllable artistic text style transfer via shape-matching GAN. In: Proceedings of International Conference Computer Vision, pp. 4442–4451 (2019)
37. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of IEEE International Conference Computer Vision and Pattern Recognition, pp. 5505–5514 (2018)
38. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of International Conference Computer Vision (2019)
39. Yu, Q., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M., Loy, C.C.: Sketch me that shoe. In: Proceedings of IEEE International Conference Computer Vision and Pattern Recognition, pp. 799–807 (2016)
40. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of International Conference Computer Vision, pp. 2242–2251 (2017)
41. Zhu, J.Y., et al.: Toward multimodal image-to-image translation. In: Advances in Neural Information Processing Systems, pp. 465–476 (2017)