Cali-Sketch: Stroke Calibration and Completion for High-Quality Face Image Generation from Human-Like Sketches

Weihao Xia[a], Yujiu Yang[b,*] and Jing-Hao Xue[c]

[a]Tsinghua University, China
[b]Tsinghua Shenzhen International Graduate School, Tsinghua University, China
[c]Department of Statistical Science, University College London, UK

ARTICLE INFO

Keywords:
Face sketch-to-photo synthesis
Image translation
Neural network
Generative adversarial network

ABSTRACT

Image generation has received increasing attention because of its wide application in security and entertainment. Sketch-based face generation brings more fun and better quality of image generation due to supervised interaction. However, when a sketch poorly aligned with the true face is given as input, existing supervised image-to-image translation methods often cannot generate acceptable photo-realistic face images. To address this problem, in this paper we propose Cali-Sketch, a human-like-sketch to photo-realistic-image generation method. Cali-Sketch explicitly models stroke calibration and image generation using two constituent networks: a Stroke Calibration Network (SCN), which calibrates strokes of facial features and enriches facial details while preserving the original intent features; and an Image Synthesis Network (ISN), which translates the calibrated and enriched sketches to photo-realistic face images. In this way, we manage to decouple a difficult cross-domain translation problem into two easier steps. Extensive experiments verify that the face photos generated by Cali-Sketch are both photo-realistic and faithful to the input sketches, compared with state-of-the-art methods.

1. Introduction

Drawing a sketch is maybe the easiest way for amateurs to describe an object or scene quickly. Compared with photographs or portraits, it does not require technical capture devices or professional painting skills. Generating photo-realistic images from free-hand sketch enables a novice to create images from their imagination, making reality a face or scene otherwise only exist in their dreams. However, the sketches drawn by non-artists are usually simple and imperfect. They are sparse, the lack of necessary details, and strokes do not precisely align with the original images or actual objects. It is hence challenging to synthesize natural and realistic images from such human-like sketches.

Recent progress on image-to-image translation [26, 70, 69, 14, 55, 7, 61] has shown that an end-to-end generative adversarial network (GAN) architecture could produce high quality results. A few of them are capable of synthesizing facial photos from sketches, but it requires sketches with precisely and strictly aligned boundaries to produce plausible results. Building such an exquisite large-scale dataset with thousands of image pairs (i.e, face photo and its corresponding sketch drawn by professional portraitists) would be quite time-consuming and expensive. It is much easier to build a dataset of face photos and their corresponding free-hand sketches drawn by amateurs. Technically, given such human-like sketches and photos, the networks proposed for cross-domain translation [26, 70, 24, 32] would learn both stroke modification and image generation simultaneously. However, the remarkable stroke and appearance differences between sketches and photos diminish the effectiveness of these networks, thus leading to unpleasant results.

There are some interactive face image modification methods under the framework of image inpainting [28, 65, 47, 35]. Given a partially-and-irregularly masked image, they refill the erased regions with strokes provided by the user as guidance. The refilled regions are consistent with input reference strokes and compatible with the whole image. Recent work [28] can obtain a realistic synthetic face photo even though the user conducts some modifications and the network tolerates minor error or mismatching. However, to generate an appropriately edited and restored result, a plausible sketch of the original image is still needed by these methods. When human-like sketches are fed into the model, the results can be unacceptable. Moreover, synthesizing an image from a total sketch is much harder than from a regionally-erased image, since in the latter case the rest edges and colors can significantly help the reconstruction.

Some methods [12, 40] consider the case of casual free-hand frontal face sketches, where the generated images do not have to strictly align with the input sketches and present more freedom in appearance. But their methods produce blurry and artifactual results. What's more, crucial components and drawing intention of the original sketches such as facial contours and hairstyles are not preserved in the synthesized images.

To address these issues, we propose a novel two-stage generative adversarial network called "Cali-Sketch", to realize face photo synthesis from human-like sketches in a unified framework. It explicitly models stroke calibration and image generation using two constituent GANs: a Stroke Calibration Network (*SCN*), which calibrates and completes strokes of facial features and enriches facial details while preserving the original intent features of the painter, and an

*Corresponding author
xiawh3@outlook.com (W. Xia); yang.yujiu@sz.tsinghua.edu.cn (Y. Yang); jinghao.xue@ucl.ac.uk (J. Xue)
ORCID(s): 0000-0003-0087-3525 (W. Xia); 0000-0002-6427-1024 (Y. Yang); 0000-0003-1174-610X (J. Xue)

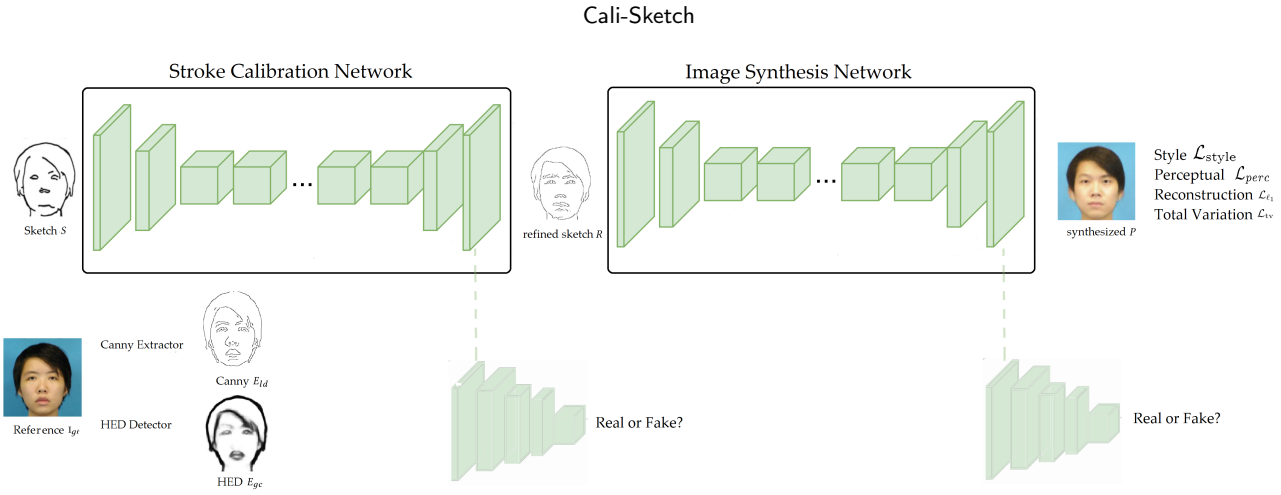

Figure 1: The overall architecture of our method. Stroke Calibration Network first calibrates unreasonable strokes and adds necessary details. The modified sketches are then fed into the Image Synthesis Network to produce photo-realistic face images.

Image Synthesis Network (*ISN*), which transfers the calibrated and completed sketches to face photos. Two GANs are first separately trained for each stage, and then trained jointly.

We focus on image generation from "human-like sketches", which is less discussed but appears more in real applications. Compared with the aforementioned methods in [26, 28, 55, 39, 38], ours doesn't necessarily desire a sketch well aligned to the original image to generate an appropriately edited and restored result while those methods might produce unacceptable results given such human-like sketches as input.

To preserve facial features and drawing intention, we propose both global contour loss and local detail loss to accomplish necessary stroke modifications and detail improvements. To eliminate artifacts, we also incorporate a perceptual loss and a reconstruction loss in the overall objective function. In this way, we manage to make the final appearance of generated images photo-realistic, while keeping the determinant attributes and drawing intention of the input sketch. Experiments confirm that face images synthesized by our proposed method are natural-looking and visually pleasant without observable artifacts.

To sum up, our key contributions are three-fold:

- We present the first two-stage human-like face sketch to photo translation. It achieves stroke calibration and image synthesis with two consecutive GANs: SCN and ISN.
- We propose SCN for necessary stroke calibration and detail completion. To preserve identity and drawing intention during the reconstruction of fine-grained face sketches, we design novel calibration loss functions. Furthermore, when given a free-hand drawn sketch, this network can act as a pre-processing modification module for other tasks using reference sketches such as interactive face image modification.
- We propose ISN for face sketch-to-image generation. The synthesized face images are both identity-consistent and appearance-realistic.

The rest of this paper is organized as follows. Section 2 provides an overview of the previous methods and related techniques. Section 3 presents the proposed Cali-Sketch method. Section 4 reports the qualitative and quantitative performance of sketch-based image synthesis experiments using the proposed method, and Section 5 summarizes and concludes the paper.

2. Related Work

2.1. Photo-Realistic Image Synthesis

Photo-realistic image synthesis methods have progressed rapidly during the last few years. The goal of image synthesis is to generate photo-realistic and faithful images from sketches or abstract semantic label maps, refer to as label-based image synthesis and sketch-based image synthesis, respectively.

Label-based image synthesis methods [55, 8, 48] synthesize image semantically from abstract label maps, such as sparse landmarks or pixel-wise segmentation maps. [55] proposes a framework for instance-level image synthesis with conditional GANs. [8] proposes a cascade framework to synthesis high-resolution images from pixel-wise labeling maps.

Facial sketch-based image synthesis approaches have been widely developed during the last few years. Those existing studies can be broadly classified into two categories: image retrieval based approaches [10, 16, 11, 51, 34] and deep learning based methods [26, 70, 12, 40, 53, 67]. The former mainly has three basic steps: retrieve, select and composite. Given a sketch plus overlaid text labels as input, Sketch2Photo [10] and Photo-Sketcher [16] automatically synthesize realistic pictures by seamlessly composing objects and backgrounds based on sketch searching and

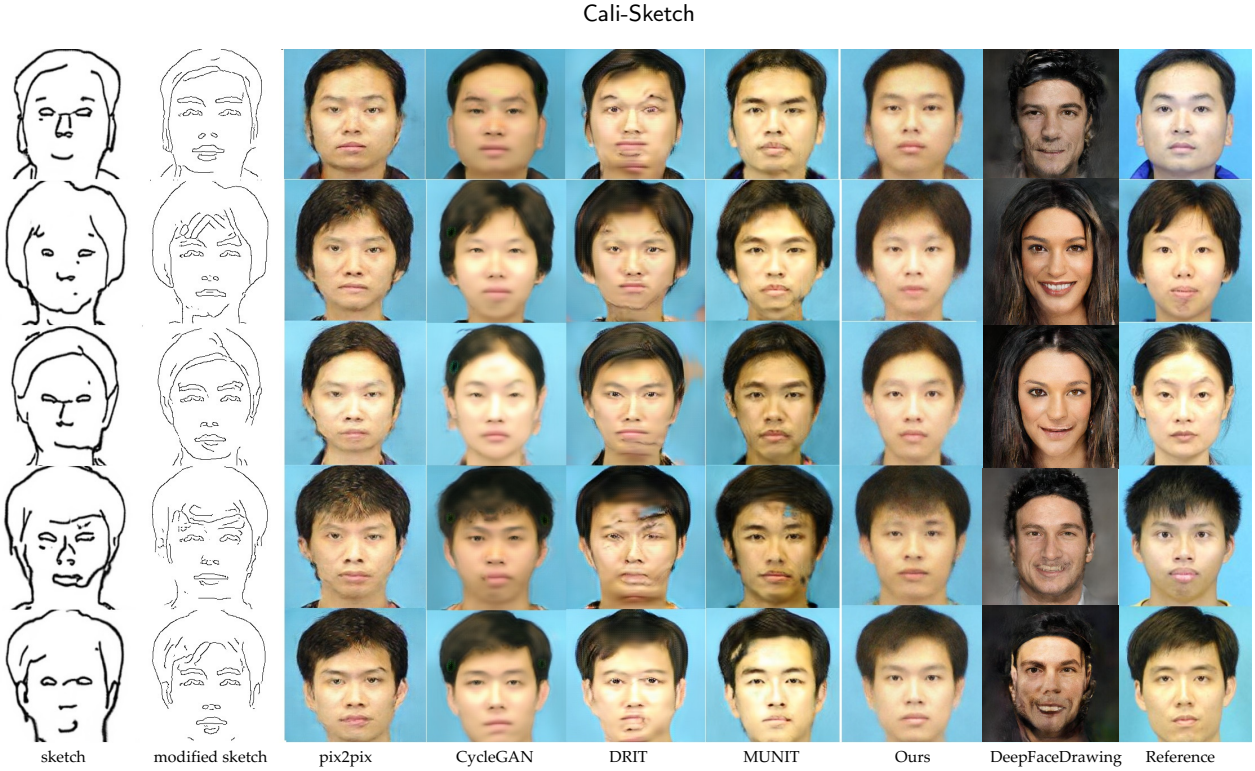

Figure 2: Qualitative comparison with baselines. We compare our methods with pix2pix [26], CycleGAN [70], DRIT [32], MUNIT [24], and a very recent sketch-to-image method DeepFaceDrawing [9]. Our approach generates more photo-realistic images. The corresponding image can be recognised easily from a batch of mixed sketches, which means crucial components and drawing intention of original sketches such as facial contours and hair styles are well-preserved in the synthesized images.

image compositing. PoseShop [11] constructs a large segmented character database for human synthesis, where people in pictures are segmented and annotated by actions and appearance attributes. Then human images are composed by feeding given sketches with text labels into the query. Those methods often suffer from heavily blurred effects and tedious inference process.

Deep learning based methods learn the mapping between sketches and photos. The pix2pix [26] translates precise edge maps to pleasing shoe pictures using conditional GANs. CycleGAN [70] proposes cycle-consistent loss to handle the paired training data limitation of pix2pix. SketchyGAN [12] synthesizes plausible images of objects from 50 classes. It aims to synthesis results both photo-realistic and faithful to the intention of given sketches. In this case, the intention was defined as generated images sharing similar poses with input sketches since it is hard to learn human intention. PhotoSketchMAN [53] generates face photos iteratively from low resolution to high resolution by multi-adversarial networks. CA-GAN [67] proposes to use pixel-wise labelling facial composition information to help face sketch-photo synthesis. Contextual-GAN [40] formulates the task of sketch-image synthesis as the joint image completion. Sketches provide contextual information for completion.

2.2. Generative Adversarial Networks

In recent years, Generative Adversarial Networks (GANs) [17] have been successfully applied in many computer vision tasks to improve the realism of generated images, such as domain adaption [46, 15], super-resolution [57, 63]. They are composed of a generator G and a discriminator D. Discriminators try to distinguish the generated fake images, while generators aim to fool discriminators from identifying real images from fake ones. The ideal solution is the Nash equilibrium where G and D couldn't improve their cost unilaterally.

Despite great success, there are still several challenges in GANs including generalization [3, 43] and training stability [2, 20]. To alleviate those problems, technologies are proposed to improve GANs. For example, Arjovsky *et al.* [2, 19] propose to minimize the Wasserstein distance between model and data distributions. Berthelot *et al.* [4] try to optimize a lower bound of the Wasserstein distances between auto-encoder loss distributions on real and fake data distributions. Mao *et al.* [42] proposes a least-squares loss for the discriminator, which implicitly minimizes Pearson χ^2 divergence, leading to stable training, high image quality and considerable diversity.

2.3. Image-to-Image Translation with GANs

General image-to-image translation methods aim to learn a mapping from the source domain to the target domain.

Isola *et al.* [26] propose a pix2pix framework trained with image pairs and achieve convincing synthetic images on many translation tasks. To handle the limitation of paired images for training, CycleGAN [70], DualGAN [64], DiscoGAN [31] present cycle consistency loss to constrain the translation between inputs and translated images. CSGAN [30] extends [70] with an additive cyclic-synthesized loss between the synthesized image of one domain and the cycled image of another domain. InstaGAN [45] incorporates instance attribute information for multi-instance transfiguration. MUNIT [24] and DRIT [32] are proposed for one-to-many diverse image translation. ComboGAN [1] also proposes a multi-component translation method without being constrained to two domains.

3. Method

3.1. Overview

Our goal is to realize face photo synthesis from a human-like sketch. Consider two data collections from different domains, $\mathbf{S} \subset I^{H\times W\times 1}$ referring to input sketch domain and $\mathbf{P} \subset I^{H\times W\times 3}$ referring to output photo domain. $I^{H\times W\times N}$ represents an image of height H, width W and channel N. Converting a face sketch from source domain $\mathbf{S}$ to an image in the target photo domain $\mathbf{P}$ can be referred to as $G : \mathbf{S} \rightarrow \mathbf{P}$. This is a typical cross-domain image translation problem but we could not directly learn the mapping by existing image-to-image translation methods. Instead, we decompose this translation into two stages: 1) Stroke Calibration Network named *SCN*, and 2) Image Synthesis Network named *ISN*. Let G_1 and D_1 be the generator and discriminator of SCN, G_2 and D_2 be the generator and discriminator of ISN, respectively. As shown in Figure 1, the input sketch $\mathbf{S}$ is first put into SCN to get the refined sketch $\mathbf{R}$ after stroke calibration and detail completion, which is then fed into ISN to generate a photo-realistic face image $\mathbf{P}$. We first train Stroke Calibration Network and Image Synthesis Network separately until the losses plateau, and then train them jointly in an end-to-end way until convergence. Qualitative comparison with baselines is demonstrated in Figure 2. Illustrations of SCN and ISN are shown in Figure 3 and 4, respectively. Training details and network architecture can be found in Section 4.2.

3.2. Stroke Calibration Network

Stroke Calibration Network aims to modify inconsequent strokes and enrich necessary details of input sketch. Let $\mathbf{S}$ be input sketches. Ground truth face photos and their edge counterparts will be denoted as $\mathbf{I}_{gt}$ and $\mathbf{E}_{gt}$. The mapping from human-like sketches $\mathbf{S}$ to the modified ones $\mathbf{R}$ can be denoted as $G_1 : \mathbf{S} \rightarrow \mathbf{R}$:

$$\mathbf{R} = G_1\left(\mathbf{S}, \mathbf{E}_{gt}\right) \tag{1}$$

where $\mathbf{E}_{gt}$ are composed of two components: global contours $\mathbf{E}_{gc}$ and local details $\mathbf{E}_{ld}$.

To modify inconsequent strokes and enrich necessary details, we introduce a novel calibration loss $\mathcal{L}_{CL}$ which consists of global contour loss and local detail loss. Global contour loss aims to modify inconsequent strokes and local detail loss enriches necessary details.

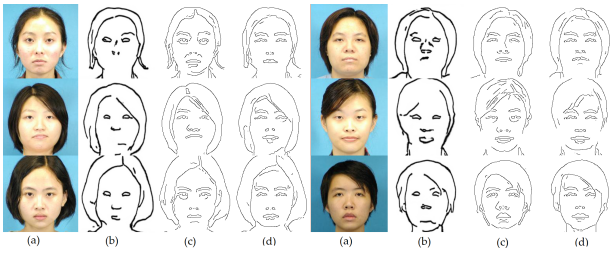

Figure 3: Illustration of Stroke Calibration Network (SCN): (a) reference; (b) input sketch; (c) Canny result; (d) modified sketch by our SCN.

We define both losses based on the feature matching loss [55]. Feature representations of real and synthesized images extracted from multiple layers of discriminator are then used to calculate the feature matching loss as

$$\mathcal{L}_{CL} = \mathbb{E}\left[\sum_{i=1}^{T}\frac{1}{N_i}\left\|D_1^{(i)}\left(\mathbf{E}_{gt}[j]\right) - D_1^{(i)}\left(\mathbf{R}\right)\right\|_1\right], \tag{2}$$

where T is the index of the final convolution layer of the discriminator, N_i is the number of elements in the i-th activation layer, $\mathbf{E}_{gt}[j], j \in \{0, 1\}$ represents global contour or local detail, and $D_1^{(i)}$ is the activation in the i-th layer of the discriminator. In our experiments, global contour and local detail are implemented by HED [60] and Canny [6] edge map, respectively. This calibration loss can stabilize training by forcing the generator to produce natural statistics at different scales [55].

For stable training, high image quality and considerable diversity as discussed in Section 2, we use the least-squares GAN [42] in our experiment. Thus, $\mathcal{L}_{\text{adv,SCN}}$ can be formulated as

$$\begin{aligned}\min_{D_1}\mathcal{L}_{\text{adv,SCN}}\left(D_1\right) =& \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\left(D_1(\boldsymbol{x}) - b\right)^2\right] + \\ & \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}\left[\left(D_1\left(G_1(\boldsymbol{z})\right) - a\right)^2\right] \\ \min_{G_1}\mathcal{L}_{\text{adv,SCN}}\left(G_1\right) =& \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}\left[\left(D_1\left(G_1(\boldsymbol{z})\right) - c\right)^2\right],\end{aligned} \tag{3}$$

where a, b, c denote the labels for fake data and real data and the value that G wants D to believe for fake data, respectively. In our experiment, x are ground truth images and z are input sketches sampled from distribution $p(\boldsymbol{z})$.

The total loss of Stroke Calibration Network combines an improved adversarial loss and calibration loss as

$$\min_{G_1}\max_{D_1}\mathcal{L}_{G_1} = \min_{G_1}\left(\max_{D_1}\left(\mathcal{L}_{\text{adv,SCN}}\right) + \lambda\mathcal{L}_{CL}\right), \tag{4}$$

where λ is regularization parameters controlling the importance of two terms. We set $\lambda = 10$ in the experiments.

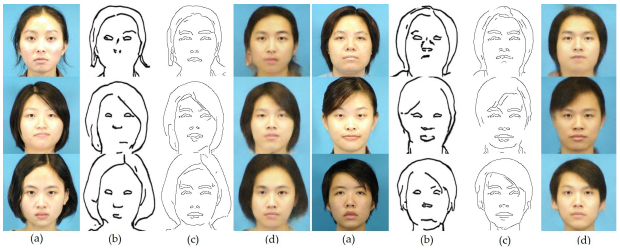

Figure 4: Illustration of Image Synthesis Network (ISN): (a) reference; (b) original input sketch; (c) input sketch modified by our SCN. (d) generated image by our ISN.

3.3. Image Synthesis Network

After stroke calibration and detail completion, the refined sketch $\mathbf{R}$ is then fed into Image Synthesis Network to generate photo-realistic face photo P. This translation process from the refined sketch $\mathbf{R}$ to the photo-realistic face image $\mathbf{P}$ can be defined as $G_2 : \mathbf{R} \rightarrow \mathbf{P}$. The output image should yield both high sketch identification similarity and favourable perceptual quality, while sharing the same resolution with the input sketch:

$$\mathbf{P} = G_2\left(\mathbf{R}, \mathbf{I}_{gt}\right) \tag{5}$$

We train this image synthesis network with a joint loss, which consists of five terms: an ℓ_1 reconstruction loss $\mathcal{L}_{\ell_1}$, adversarial loss $\mathcal{L}_{adv,2}$, perceptual loss $\mathcal{L}_{\text{percep}}$, style loss $\mathcal{L}_{\text{style}}$ and total variation loss $\mathcal{L}_{\text{tv}}$:

$$\mathcal{L}_{G_2} = \lambda_1\mathcal{L}_{\ell_1} + \lambda_2\mathcal{L}_{\text{adv,ISN}} + \lambda_3\mathcal{L}_{\text{percep}} + \lambda_4\mathcal{L}_{\text{style}} + \mathcal{L}_{\text{tv}} \tag{6}$$

Reconstruction loss $\mathcal{L}_{\ell_1}$ minimizes the differences between reference and generated images:

$$\mathcal{L}_{\ell_1} = \mathbb{E}\Big[\sum_i \frac{1}{N_i}\Big\|\mathbf{I}_{gt} - \mathbf{P}\Big\|_1\Big] \tag{7}$$

Perceptual loss $\mathcal{L}_{\text{percep}}$ is proposed by Johnson *et al.* [29] based on perceptual similarity. It is originally defined as the distance between two activated features of a pre-trained deep neural network. Here we adopt a more effective perceptual loss which uses features before activation layers [57]. These features are more dense and thus provide relatively stronger supervision, leading to better performance:

$$\mathcal{L}_{\text{percep}} = \mathbb{E}[\sum_i \frac{1}{N_i}\|\phi_i(\mathbf{I}_{gt}) - \phi_i(\mathbf{P})\|_1], \tag{8}$$

where ϕ_i donates the feature maps before activation of the VGG-19 network pre-trained for image classification.

Style loss $\mathcal{L}_{\text{style}}$ is adopted in the same form as in the original work [29], which aims to measure differences between covariance of activation features:

$$\mathcal{L}_{\text{style}} = \mathbb{E}_j[\|G_j^{\phi}(\mathbf{I}_{gt}) - G_j^{\phi}(\mathbf{P})\|_1], \tag{9}$$

where G_j^{ϕ} represents the Gram matrix constructed from feature maps ϕ_j.

Total variation loss is based on the principle that images with unrestrained and possibly spurious detail have high total variation. According to this, reducing the total variation of an image subject to it being a close match to the original image, removes unwanted noises while enforcing spatial smoothness and preserving important details such as edges. It is defined on the basis of the absolute gradient of generated images:

$$\mathcal{L}_{\text{tv}} = \|\nabla_x\mathbf{P} - \nabla_y\mathbf{P}\|_1. \tag{10}$$

For experiments, we use $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.1$, and $\lambda_4 = 200$.

4. Experiments

4.1. Training Data

Appropriate and adequate training data is important for network performance. Since it is infeasible to collect large-scale paired images and sketches, most existing free-hand sketch based image synthesis methods generate sketches automatically from images.

To exhibit different styles of free-hand sketches and to improve the network generality, we augment training data by adopting multiple different styles of input sketches. Specifically, we generate four different free-hand sketch styles in total. We use the XDoG edge detector [58] and Photocopy effect in Photoshop to generate two styles. To better resemble hand-drawn sketches, we simplified the edge images using [50] as in [40]. We also use photo-sketch [33] to generate the desired face sketches. This recent method generates imperfect alignment contour sketches of input images. The human-like sketches should be sparse and contain these wrong edges. That's why the Canny algorithm [6] shouldn't be chosen to get input sketches. Those edges generated by Canny are solid and well-aligned with input images. To show the effectiveness and efficiency of our approach, the CUHK Face Sketch Database [56] is used in our experiment for its appropriateness and popularity. We use its $256 \times 256 \times 3$ resized and cropped version.

Figure 5 are illustration of well-drawn sketches from [56]. These sketches are drawn by the artist. Compared with human-like sketches in Figure 6, the well-depicted sketches capture the most distinctive characteristics of human faces and are faithful to the original face images. We often can easily recognize a person from the corresponding sketch. The free-hand sketches are often sparse, deformed, the lack of necessary strikes or details and lines do not precisely align to the real face images, sometimes even the lack of necessary lines in the area of mouth or jaw as illustrated in Figure 6.

The ground truth sketches for Stroke Calibration Network are generated using Canny algorithm [6] and Holistically-nested Edge Detection (HED) edge detector [60]. Specifically, we extract HED from images after histogram equalization to avoid the interference of light. Thus, we generate a desired new dataset consisted of high-quality face photos and corresponding human-like face sketches, Canny together with HED edges.

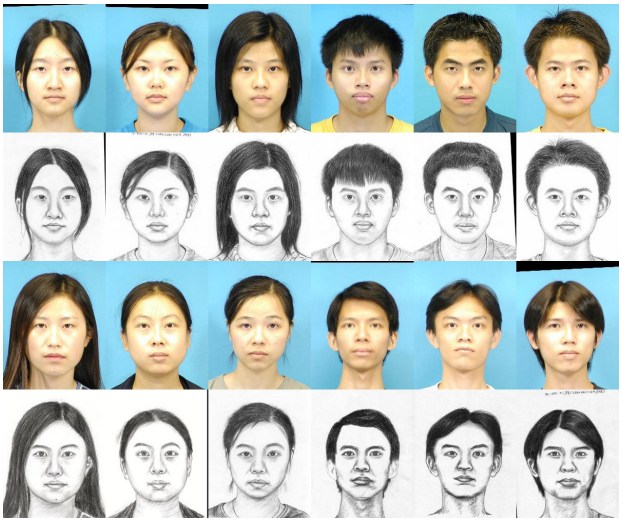

Figure 5: Illustration of well-drawn sketches from [56]. Best viewed in color.

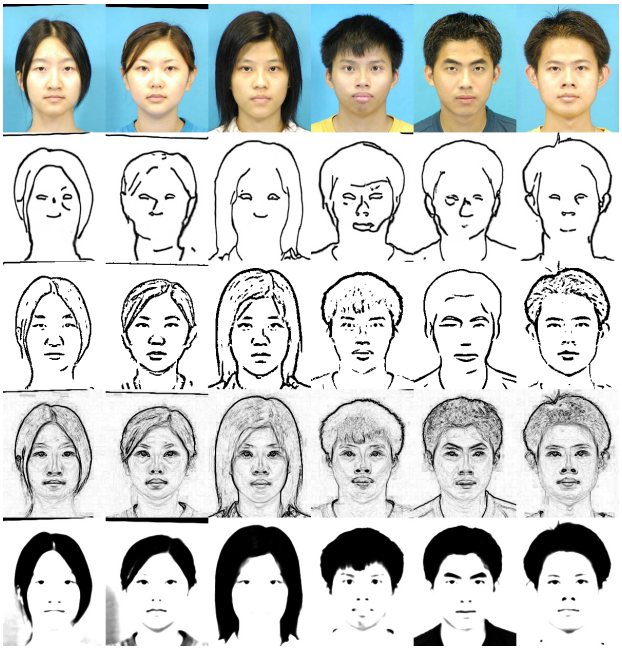

Figure 6: Illustration of four different free-hand sketch styles: photo-sketch [33], XDoG [50], Photocopy of Photoshop and FDoG [58].

4.2. Experiment Settings

Network architecture. Inspired by recent image translation studies, our generators follow an encoder-decoder architecture similar to the method proposed by Johnson *et al.* [29]. Each of the generators consists of two down-sampling encode layers, followed by eight residual blocks[21] and two up-sampling decoders. Skip connections are added to concatenate previous layers with the same spatial resolution. We replace regular convolutions in the residual blocks with dilated convolutions with dilation factor two to obtain large receptive fields. Our discriminators are based on the SN-PatchGAN [66] architecture, which determines whether or not overlapping image patches of a certain size are real. Spectral normalization [44] is introduced for rapid and stable training and helps produce high-quality results.

Notice that here we did not deliberately design the structure of the Image Synthesis Network (SCN). In fact, we adopt a quite simple structure of SCN to show that it is easy to generate satisfactory results from the calibrated sketches even the sktech2image synthesis network is not deliberately designed. For more details about the indispensability of Stroke Calibration Network and scalability of Image Synthesis Network, refer to Section 4.5.2 and 4.5.3 respectively.

Training strategy. The training strategy is demonstrated in Algorithm 1. Forward and backward in Algorithm 1 represent forward propagation and back propagation respectively. The forward process includes steps of passing the input through the network layers and calculating the actual output and losses of the model. The backward process back-propagates errors and updates the weights of the network. We refer corresponding operations to as forward and backward for simplicity and emphasize that our method is an end-to-end method with three-stage training. N_1, N_2, N_3 are iteration numbers which are large enough to guarantee convergence. Firstly, we train our Stroke Calibration Network G_1 using the Canny and HED edges as supervision with a 10^{-4} learning rate. Meanwhile, we train Image Synthesis Network G_2 using Canny $\odot$ HED as input refined sketches and ground truth face images as supervision with the same 10^{-4} learning rate. Here, $\odot$ denotes the Hadamard product. We then decrease the learning rate to 10^{-5} and jointly train both G_1 and G_2 in an end-to-end way until convergence. Discriminators are trained with a learning rate of one-tenth of the generators' according to different training stages. Both networks are trained with resized 256×256 images with a batch of 8.

Evaluation metrics. For our task of face image synthesis from human-like sketches, we use two kinds of evaluation metrics: similarity metrics and perceptual scores. We apply the widely used full reference image quality assessment metrics such as PSNR, SSIM as similarity metrics. Given two images $I, I' \subset I^{H \times W \times C}$, the peak signal-to-noise ratio (PSNR) are defined as

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{\mathrm{L}^2}{\mathrm{MSE}} \right), \tag{11}$$

where L is usually 255, $\| \cdot \|_F^2$ is the Frobenius norm and $\mathrm{MSE} = \frac{1}{HWC} \| I - I' \|_F^2$. The structural similarity index (SSIM) is defined as

$$\mathrm{SSIM}(I, I') = \frac{2\mu_I \mu_{I'} + k_1}{\mu_I^2 + \mu_{I'}^2 + k_1} \cdot \frac{\sigma_{II'} + k_2}{\sigma_I^2 + \sigma_{I'}^2 + k_2}, \tag{12}$$

where μ_I and σ_I^2 is the mean and variance of $I, \sigma_{I\hat{I}}$ is the covariance between I and $\hat{I}$, and k_1 and k_2 are constant relaxation terms. A highest score indicates a more structurally similar face for a given sketch.

Algorithm 1: TRAINING STRATEGY

Stage 1: SCN training
Input: **S**, free-hand sketch.
Output: **R**, Refined sketch
while $n \leq N_1$ **do**
  $\mathbf{R}, \mathcal{L}_{G_1}, \mathcal{L}_{D_1} = G_1.\text{forward}((\mathbf{S}, \mathbf{E}_{gt}))$
  $G_1.\text{backward}$
Stage 2: ISN training
Input: **R**, Canny⊙ HED.
Output: **P**, Generated face image.
while $n \leq N_2$ **do**
  $\mathbf{P}, \mathcal{L}_{G_2}, \mathcal{L}_{D_2} = G_2.\text{forward}((\mathbf{R}, \mathbf{I}_{gt}))$
  $G_2.\text{backward}$
Stage 3: Joint training
Input: **S**, free-hand sketch.
Output: **P**, Generated face image.
while $n \leq N_3$ **do**
  $\mathbf{R}, \mathcal{L}_{G_1}, \mathcal{L}_{D_1} = G_1.\text{forward}((\mathbf{S}, \mathbf{E}_{gt}))$
  $\mathbf{P}, \mathcal{L}_{G_2}, \mathcal{L}_{D_2}.\text{forward}((\mathbf{R}, \mathbf{I}_{gt}))$
  $G_1.\text{backward}$
  $G_2.\text{backward}$

For perceptual scores, we use Fréchet Inception distance (FID) [23]. The FID is defined using the Fréchet distance between two multivariate Gaussians:

$$\text{FID} = \left\|\mu_r - \mu_g\right\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right), \quad (13)$$

where $X_r \sim \mathcal{N}\left(\mu_r, \Sigma_r\right)$ and $X_g \sim \mathcal{N}\left(\mu_g, \Sigma_g\right)$ are the 2048-dimensional activations of the Inception-v3 pool3 layer for real and generated samples respectively The lowest FID means it achieves the most perceptual results.

4.3. Baselines

We perform the evaluation on the following baseline methods:

Pix2pix [26] is an early work of image-to-image translation. It achieves good photo results on edge-to-photo generation, and the models trained on automatically detected edges can generalize to human drawings.

CycleGAN [70] achieves unsupervised image-to-image translation via cycle-consistent loss.

DRIT [32] is a recent work, which realizes diverse image-to-image translation via disentangled content and attribute representations of different domains. Experiment on the edge-to-shoes dataset shows it can produce both realistic and diverse images.

MUNIT [24] is the state-of-the-art unsupervised multi-domain image-to-image translation framework. It achieves quality and diversity comparable to the state-of-the-art supervised algorithms on the task of edge-to-shoes/handbags.

DeepfaceDrawing [9] is a recent state-of-the-art sketch-to-image framework. The key idea of their method is to implicitly model the shape space of plausible face images and synthesize a face image in this space to approximate an input sketch in a local-to-global way.

For fair evaluation, all baselines are retrained with the sketch-image pairs except for DeepFaceDrawing [9], which we use their online demo [1] since the training scripts are not available. We do not compare with the recent human-drawn sketch to image method [62] since its implementation is not publicly available.

4.4. Comparison Against Baselines

Qualitative evaluation. Qualitative comparison with baselines are demonstrated in Figure 2. The results produced by pix2pix [26] all have obvious artifacts. All facial features suffer from shape distortion to a degree, especially the facial and ear contours on the first and fifth rows. CycleGAN [70] produces the most similar face with the reference, but its results are blurry and unpleasing. There are two or more visible spots in the area of hair. The contours of face images generated by DRIT [32] are aligned with their lines of the input sketches, which notably deteriorate the image quality. MUNIT [24] could produce relatively visually realistic and qualitatively consistent results. however, they are more like oil paintings rather than photos.

Compared with baseline methods, our approach generates high-quality images. The generated human face images are more photo-realistic. The corresponding image can be recognized easily from a batch of mixed sketches, which means crucial components and drawing intention of original sketches like facial contours, hairstyles are well-preserved in the synthesized images.

Quantitative comparison. Quantitative evaluation with baselines is shown in Table 1. For PSNR and SSIM, CycleGAN [70] achieves the highest structural similarity, and our method ranks the second. For the task of sketch-to-image generation, the similarity is not that important, since there are no corresponding real face images as reference for most free-hand drawn sketches. What really matters is whether generated images are photo-realistic or not.

Fréchet Inception Distance (FID) is calculated by computing the Fréchet distance between two Gaussians of feature representations extracted from the pre-trained inception network [52]. It is not only a measure of similarity between two datasets of images, but also shown to correlate well with the human visual judgement of image quality. Due to the above advantages, FID [23] is most often used to evaluate the quality of images generated by Generative Adversarial Networks. As shown in Table 1, our method achieves the lowest FID score, which means that our method produces the best results in both perceptual judgement and high-level similarity.

4.5. Ablation Studies

4.5.1. The choice of contour and detail

There are many choices of contour and detail for our methods such as edges, boundaries and contours. These are a few differences between them. Edge maps are precisely aligned

[1] http://geometrylearning.com/DeepFaceDrawing/

Table 1
Perfomance as PSNR, SSIM and FID on the CUHK dataset. The **best** and second best results are highlighted in each column. For details refer to Section 4.4.

Method	PSNR	SSIM	FID
pix2pix [26]	18.83	0.7554	76.90
CycleGAN [70]	**24.21**	**0.8508**	80.17
MUNIT [24]	17.23	0.7515	78.57
DRIT [32]	16.14	0.7047	109.83
DeepFaceDrawing [9]	13.21	0.3751	97.36
Ours	20.25	0.8006	**58.43**

to object boundaries, and they usually contain more information about details and backgrounds. Boundaries pay more attention to external lines. Contours contain object boundaries, salient internal and background edges. Contours can be obtained by the boundary contour edge extractors like HED [60], COB [41], RCF [36, 37], or similar to pix2pixHD [55], simplified from the face parsing semantic labels. For a face image, contours are more like facial feature boundaries.

Since sketches are the approximate outline of the objects with spatial transformations and deformed strokes [27], we need to modify its strokes and add more details before synthesis. Contours and edges are respectively responsible for stroke calibration and detail completion. We will illustrate the reasons in the next part. In our experiment, we choose HED as global contour and Canny as ground truth local detail for simplicity.

4.5.2. The impact of Stroke Calibration Network

We have tested directly applying the pix2pix to generate face images from human-like sketches, but found the training unstable and the quality of results unsatisfactory. The original sketches are deformed and sometimes lack of necessary lines in the area of mouth or jaw, as shown in Figure 3 and Figure 7. It inspires us to modify strokes and add essential details before image synthesis. Edges like the Canny detector can act as ground truth for the training of this process. The refined sketches are more visually favourable and consistent with the original identity, as shown in the third column of Figure 7.

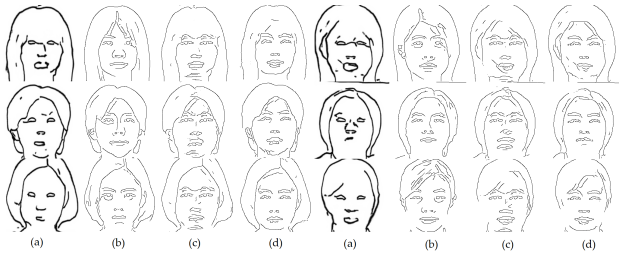

Figure 7: The impact of Stroke Calibration Network and with or without global contour. (a) input sketch. (b) canny edge. (c) result with only local detail loss. (d) result with both local detail loss and global contour loss.

However, it also demonstrates that only being supervised by Canny is not enough. It results in unwanted strokes in the area of the eyebrow mouth or jaw, and even changes the shape of the eyes. Stroke calibration should be superior to detail enrichment. We want the Stroke Calibration Network to modify strokes without changing holistic properties like facial contours and hairstyles. So we add both the contours as the global constraint and the edges as a local constraint. As shown in the fourth column of Figure 7, it calibrates unreasonable strokes and preserves original properties.

Table 2
Quantitative performance of our Stroke Calibration Network trained on only local detail loss(detail only), global contour loss(contour only) and local detail loss together with global contour loss(detail and contour).

Components	Precision	Recall
detail only	0.1559	0.1475
contour only	0.1713	0.1564
detail and contour	**0.9962**	**0.4772**

Table 2 shows the accuracy of our Stroke Calibration Network. We measure precision and recall with Canny for ablation study of local detail loss and global contour loss. For each setting, we convert the refined sketch and corresponding Canny to binary with a constant threshold value (*i.e.*, each pixel is either zero or one). Precision means how many ones in the refined sketch are actual ones in the ground truth Canny, and recall means how many ones in the ground truth Canny are contained in the refined sketch. The high precision and relatively low recall are in line with expectations. The original purpose of the Stroke Calibration Network is to modify unreasonable strokes and add essential details. The low accuracy of using local detail loss only is consistent with results in Figure 7. Since HED and Canny are different, it is not surprising that the accuracy of using global contour loss only is low.

4.5.3. The scalability of Image Synthesis Network

Notice that in Section 3.3 we didn't deliberately design the structure of the Image Synthesis Network (SCN). In fact, we adopt a quite simple structure of SCN to show that it is easy to generate satisfactory results from the calibrated sketches even the sktech2image synthesis network is not deliberately designed. Since there are many methods [28, 65] for "well-drawn" sktech2image problem, we argue that such stroke calibration is indispensable for these methods to be well applied in some real applications, such as cultural relics or digital sketch generation for suspects, to produce realistic images. Therefore, it is a useful application and a new solution to synthesize a high-quality image from human-like sketches. The results in Section 4.4 have shown that our proposed stroke calibration network is a simple yet effective. The calibrated sketches can be directly fed into other existing "well-drawn" sktech2image methods [28, 65] to produce more diverse and more photo-realistic results. Our two-stage Algorithm 1 provides end-to-end scalability for improving SCN by designing novel architecture or combining with ex-

Table 3
The scalability of Image Synthesis Network. We compare results generated by different structures of SCN on the CUHK dataset using PSNR, SSIM and FID as metrics. For details refer to Section 4.4 and Section 4.5.3.

Method	PSNR	SSIM	FID
Original	20.25	0.8006	58.43
Improved-1	20.34	0.8092	57.09
Improved-2	21.09	0.8137	55.12

isting "well-drawn" sktech2image methods.

For example, we can improve SCN by simply doubling the numbers of residual blocks (refer to as Improved-1). Or building our generator based on U-Net and using Masked Residual Unit (MRU) module proposed in [12] (refer to as Improved-2). MRU is shown to be more effective than ResNet, Cascaded Refinement Network (CRN) or DCGAN structure in image synthesis task according to [12]. We compare images generated by different structures of SCN on the CUHK dataset using PSNR, SSIM and FID as metrics. The results are shown in Table 3.

5. Conclusion and Discussion

We propose a human-like sketch based face image synthesis method named *Cali-Sketch*. Our method can generate pleasing results even when the input sketches are not plausible. To achieve this, we introduce a two-stage sketch-to-image translation method consisting of two GANs. Stroke Calibration Network first calibrates unreasonable strokes and adds necessary details. The refined sketches are then fed into the Image Synthesis Network to produce photo-realistic face images. Given human-like sketches, Cali-Sketch can generate identity-consistent and appearance-realistic face images. Experimental results show the effectiveness and efficiency of the proposed Cali-Sketch, showing superior performance than the state-of-the-art methods.

5.1. Analysis for Introducing the Refined Sketch R

In the first place, we try to directly learn the mapping from the sketch to the image using some state-of-the-art methods, *e.g.*, pix2pix [26], CycleGAN [70] and their variants. However, these existing one-stage methods are unable to generate reasonable and high-quality images directly from the "badly-drawn sketches". The reason is that these one-stage methods are designed for and trained on sketches with precisely and strictly aligned boundaries. They necessarily desire a plausible sketch referring to the original image. When it came to image generation from "badly-drawn sketches", which is less discussed but appears more in real applications, these methods might produce unacceptable results given such badly-drawn sketches as input, for example, the restoration results from the open-source demo of SC-FEGAN [28] when the given sketches are badly-drawn, as shown in Figure 8. Thus, we divide *image generation from human-like sketches* into a two-stage process: stroke calibration and image generation. Stroke calibration is solely focused on hallucinating edges in the missing regions. The image synthesis network uses the hallucinated edges to generate the final results.

There are two major benefits by introducing refined sketch R. On the one hand, since there are many methods for the "well-drawn" sketch-to-image problem, we argue that such a stroke calibration mechanism is indispensable for the restoration in some real applications like cultural relics or the task of creating digital sketches of suspects, and plays an important role in the final production of realistic images. In fact, we adopt a simple structure of SCN to show that it is easy to generate a satisfactory result from the calibrated sketches even the network is not deliberately designed. The results show that our proposed stroke calibration network is simple yet effective. On the other hand, since our stroke calibration adopts the popular HED [60] and Canny [6], the calibrated sketches can be directly fed into other existing "well-drawn" sketch-to-image methods [28, 65, 56, 26] to produce more diverse and more photo-realistic results.

5.2. Extended Application

Our Stroke Calibration Network can act as a pre-processing module for real-world sketches. For interactive face image manipulation like [28], a plausible input sketch is necessary. When free-hand drawn sketches are directly fed into those models, the results may be unacceptable. In this case, our Stroke Calibration Network can also act as pre-processing modification module. [28] is a recent facial image editing method. Users draw sketch S and color as guidance on incomplete image $I \odot M$ erased by mask M. To show the effectiveness and efficiency of our approach, in this case, we first directly use original human-like sketch S as input sketch for [28] to get an edited image. Then we feed the refined sketch R pre-processed by our Stroke Calibration Network to produce another edited image. As demonstrated in Figure 8, when the input sketch is sparse and contains wrong strokes and directly fed into [28], the generated facial features are distorted and deformed. Our Stroke Calibration Network can calibrate unreasonable strokes and add necessary details. When this refined sketch is fed into [28], the result is improved significantly.

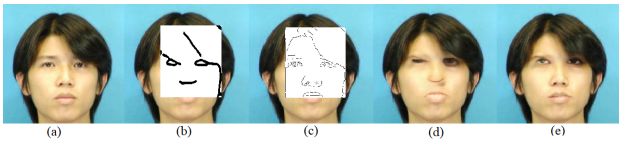

Figure 8: Stroke Calibration Network as a pre-processing module of [28] for real-world sketches. (a) original image. (b) masked image and input sketch. (c) masked image and modified sketch. (d) editing result of original sketch. (e) editing result of modified sketch.

Our method also show its potential for sketch-based object search [5, 22] and image retrieval [18, 25, 13]. Various works have been proposed to efficiently support automatic annotation of multimedia contents and help content-based retrieval, but obtaining precise image samples suffic-

ing the user specification may not be always handy. In such cases, the sketch can be an alternative solution to initialize the search, *i.e.*, sketch based image retrieval [5, 68]. Our method can help complete necessary object information critical for a reliable search performance.

5.3. Limitation and Future Work

Compared with image inpainting [28, 65] or image-to-sketch synthesis [49, 54, 39], generating photo-realistic image from human-like sketch is more challenging since there is less information in sketches. Thus, we temporarily experiment on frontal faces without large rotation and translation. The dataset limitations provide strong motivation for future work to improve performance by expanding the datasets into faces with various angles or expressions, and further into all classes, *e.g.*, the Google QuickDraw Dataset. In addition, the category of a sketch is also critical for image generation. Sketches are far from being complete in terms of the object information that would be transformed into a totally different object during generation. For example, as illustrated in Figure 9, if a user is intent on generating a pyramid image by simply drawing a 'triangle', it is not sufficiently discriminative to uniquely resemble the pyramids. Thus, incorporating category information or textual descriptions [59] of human-like sketches is critical. We will develop our Cali-Sketch into a drawing assistance that creates photographic self-portraits or user's favorite cartoon characters.

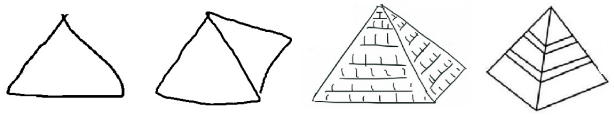

Figure 9: Sketch samples of 'triangle'. It is not sufficiently discriminative to uniquely resemble the pyramids by simply drawing a 'triangle'. Incorporating category information is critical for image generation from human-like sketches.

6. Acknowledgements

This work was partly supported by the Major Research Plan of National Natural Science Foundation of China (Grant No. 61991451), and Shenzhen special fund for the strategic development of emerging industries (Grant No. ZDYBH201900000002).

References

[1] Anoosheh, A., Agustsson, E., Timofte, R., Van Gool, L., 2018. Combogan: Unrestrained scalability for image domain translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition Workshops, pp. 783–790.

[2] Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks, in: Proceedings of International Conference on Machine Learning, pp. 214–223.

[3] Arora, S., Ge, R., Liang, Y., Ma, T., Zhang, Y., 2017. Generalization and equilibrium in generative adversarial nets, in: Proceedings of International Conference on Machine Learning, pp. 224–232.

[4] Berthelot, D., Schumm, T., Metz, L., 2017. Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717 .

[5] Bhattacharjee, S.D., Yuan, J., Huang, Y., Meng, J., Duan, L., 2018. Query adaptive multiview object instance search and localization using sketches. IEEE Transactions on Multimedia 20, 2761–2773.

[6] Canny, J., 1986. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence , 679–698.

[7] Chen, L., Wu, L., Hu, Z., Wang, M., 2019. Quality-aware unpaired image-to-image translation. IEEE Transactions on Multimedia 21, 2664–2674.

[8] Chen, Q., Koltun, V., 2017. Photographic image synthesis with cascaded refinement networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1520–1529.

[9] Chen, S.Y., Su, W., Gao, L., Xia, S., Fu, H., 2020. DeepFaceDrawing: Deep generation of face images from sketches. TOG 39, 72:1–72:16.

[10] Chen, T., Cheng, M.M., Tan, P., Shamir, A., Hu, S.M., 2009. Sketch2Photo: Internet image montage. ACM transactions on graphics 28, 1–10.

[11] Chen, T., Tan, P., Ma, L.Q., Cheng, M.M., Shamir, A., Hu, S.M., 2013. Poseshop: Human image database construction and personalized content synthesis. IEEE Transactions Visualization and Computer Graphics 19, 824–837.

[12] Chen, W., Hays, J., 2018. SketchyGAN: Towards diverse and realistic sketch to image synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9416–9425.

[13] Choi, J., Cho, H., Song, J., Yoon, S.M., 2019. Sketchhelper: Real-time stroke guidance for freehand sketch retrieval. IEEE Transactions on Multimedia 21, 2083–2092.

[14] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J., 2018. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797.

[15] Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J., 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 994–1003.

[16] Eitz, M., Richter, R., Hildebrand, K., Boubekeur, T., Alexa, M., 2011. Photosketcher: interactive sketch-based image synthesis. Computer Graphics and Applications 31, 56–66.

[17] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in Neural Information Processing Systems, pp. 2672–2680.

[18] Grigorova, A., De Natale, F.G.B., Dagli, C., Huang, T.S., 2007. Content-based image retrieval by feature adaptation and relevance feedback. IEEE Transactions on Multimedia 9, 1183–1192.

[19] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A., 2017a. Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems, pp. 5767–5777.

[20] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017b. Improved training of Wasserstein GANs, in: Advances in Neural Information Processing Systems, pp. 5767–5777.

[21] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

[22] He, S., Zhou, Z., Farhat, F., Wang, J.Z., 2018. Discovering triangles in portraits for supporting photographic creation. IEEE Transactions on Multimedia 20, 496–508.

[23] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Proc. Adv. Neural Inf. Process. Syst., pp. 6626–6637.

[24] Huang, X., Liu, M.Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image-to-image translation, in: Proceedings of the European Conference on Computer Vision, pp. 172–189.

[25] Ioannakis, G., Koutsoudis, A., Pratikakis, I., Chamzas, C., 2018.

Retrieval¡ªan online performance evaluation tool for information retrieval methods. IEEE Transactions on Multimedia 20, 119–127.

[26] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5967–5976.

[27] Jing, Y., Liu, Y., Yang, Y., Feng, Z., Yu, Y., Tao, D., Song, M., 2018. Stroke controllable fast style transfer with adaptive receptive fields, in: European Conference on Computer Vision, pp. 238–254.

[28] Jo, Y., Park, J., 2019. SC-FEGAN: Face editing generative adversarial network with user's sketch and color, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1745–1753.

[29] Johnson, J., Alahi, A., Li, F.F., 2016. Perceptual losses for real-time style transfer and super-resolution, in: Proceedings of the European Conference on Computer Vision, pp. 694–711.

[30] Kancharagunta, K.B., Dubey, S.R., 2019. Csgan: Cyclic-synthesized generative adversarial networks for image-to-image transformation. arXiv preprint arXiv:1901.03554 .

[31] Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J., 2017. Learning to discover cross-domain relations with generative adversarial networks, in: Proceedings of International Conference on Machine Learning, pp. 1857–1865.

[32] Lee, H.Y., Tseng, H.Y., Mao, Q., Huang, J.B., Lu, Y.D., Singh, M., Yang, M.H., 2020. Drit++: Diverse image-to-image translation via disentangled representations. International Journal of Computer Vision , 1–16.

[33] Li, M., Lin, Z., M˘ ech, R., , Yumer, E., Ramanan, D., 2019. Photo-sketching: Inferring contour drawings from images. IEEE Winter International Conference on on Applications of Computer Vision .

[34] Liang, Y., Song, M., Xie, L., Bu, J., Chen, C., 2012. Face sketch-to-photo synthesis from simple line drawing, in: Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, IEEE. pp. 1–5.

[35] Liu, J., Yang, S., Fang, Y., Guo, Z., 2018. Structure-guided image inpainting using homography transformation. IEEE Transactions on Multimedia 20, 3252–3265.

[36] Liu, Y., Cheng, M., Hu, X., Bian, J., Zhang, L., Bai, X., Tang, J., 2019a. Richer convolutional features for edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 1939–1946.

[37] Liu, Y., Cheng, M.M., Hu, X., Bian, J.W., Zhang, L., Bai, X., Tang, J., 2019b. Richer convolutional features for edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence .

[38] Liu, Y., Qin, Z., Wan, T., Luo, Z., 2018. Auto-painter: Cartoon image generation from sketch by using conditional wasserstein generative adversarial networks. Neurocomputing 311, 78–87.

[39] Lu, D., Chen, Z., Wu, Q.J., Zhang, X., 2019. FCN based preprocessing for exemplar-based face sketch synthesis. Neurocomputing 365, 113–124.

[40] Lu, Y., Wu, S., Tai, Y.W., Tang, C.K., 2018. Image generation from sketch constraint using contextual gan, in: Proceedings of the European Conference on Computer Vision, pp. 205–220.

[41] Maninis, K., Pont-Tuset, J., Arbeláez, P., Van Gool, L., 2018. Convolutional oriented boundaries: From image segmentation to high-level tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 819–833.

[42] Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802.

[43] Mescheder, L., Nowozin, S., Geiger, A., 2017. The numerics of gans, in: Advances in Neural Information Processing Systems, pp. 1825–1835.

[44] Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y., 2018. Spectral normalization for generative adversarial networks, in: Proceedings of International Conference on Learning Representations, p. 1.

[45] Mo, S., Cho, M., Shin, J., 2019. Instagan: Instance-aware image-to-image translation, in: Proceedings of International Conference on Learning Representations, p. 1.

[46] Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K., 2018. Image to image translation for domain adaptation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4500–4509.

[47] Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M., 2019. Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 .

[48] Qi, X., Chen, Q., Jia, J., Koltun, V., 2018. Semi-parametric image synthesis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8808–8816.

[49] Qi, Y., Guo, J., Song, Y.Z., Xiang, T., Zhang, H., Tan, Z.H., 2015. Im2Sketch: Sketch generation by unconflicted perceptual grouping. Neurocomputing 165, 338–349.

[50] Simo-Serra, E., Iizuka, S., Sasaki, K., Ishikawa, H., 2016. Learning to simplify: fully convolutional networks for rough sketch cleanup. ACM transactions on graphics 35, 121:1–121:11.

[51] Song, M., Chen, C., Bu, J., Sha, T., 2012. Image-based facial sketch-to-photo synthesis via online coupled dictionary learning. Information Sciences 193, 233–246.

[52] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826.

[53] Wang, L., Sindagi, V., Patel, V., 2018a. High-quality facial photo-sketch synthesis using multi-adversarial networks, in: IEEE Conference Automatic Face & Gesture Recognition, pp. 83–90.

[54] Wang, N., Zhu, M., Li, J., Song, B., Li, Z., 2017. Data-driven vs. model-driven: Fast face sketch synthesis. Neurocomputing 257, 214–221.

[55] Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B., 2018b. High-resolution image synthesis and semantic manipulation with conditional GANs, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8798–8807.

[56] Wang, X., Tang, X., 2009. Face photo-sketch synthesis and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 1955–1967.

[57] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C., 2018c. ESRGAN: Enhanced super-resolution generative adversarial networks, in: Proceedings of the European Conference on Computer Vision, pp. 63–79.

[58] Winnemöller, H., Kyprianidis, J.E., Olsen, S.C., 2012. Xdog: An extended difference-of-gaussians compendium including advanced image stylization. Computers & Graphics 36, 740–753.

[59] Xia, W., Yang, Y., Xue, J.H., Wu, B., 2020. Tedigan: Text-guided diverse image generation and manipulation. arXiv preprint arXiv:2012.03308 .

[60] Xie, S., Tu, Z., 2015. Holistically-nested edge detection, in: Proceedings of the IEEE international conference on computer vision, pp. 1395–1403.

[61] Xu, W., Keshmiri, S., Wang, G., 2019. Adversarially approximated autoencoder for image generation and manipulation. IEEE Transactions on Multimedia 21, 2387–2396.

[62] Yang, S., Wang, Z., Liu, J., Guo, Z., 2020. Deep plastic surgery: Robust and controllable image editing with human-drawn sketches. arXiv preprint arXiv:2001.02890 .

[63] Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.H., Liao, Q., 2019. Deep learning for single image super-resolution: A brief review. IEEE Transactions on Multimedia 21, 3106–3121.

[64] Yi, Z., Zhang, H., Tan, P., Gong, M., 2017. Dualgan: Unsupervised dual learning for image-to-image translation, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2849–2857.

[65] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2018. Generative image inpainting with contextual attention, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5505–5514.

[66] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2019. Free-form image inpainting with gated convolution, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4471–4480.

[67] Yu, J., Xu, X., Gao, F., Shi, S., Wang, M., Tao, D., Huang, Q., 2020.

Toward realistic face photo-sketch synthesis via composition-aided GANs. IEEE Transactions on Cybernetics .

[68] Zhang, Y., Qian, X., Tan, X., Han, J., Tang, Y., 2016. Sketch-based image retrieval by salient contour reinforcement. IEEE Transactions on Multimedia 18, 1604–1615.

[69] Zhou, Y.F., Jiang, R.H., Wu, X., He, J.Y., Weng, S., Peng, Q., 2019. BranchGAN: Unsupervised mutual image-to-image transfer with a single encoder and dual decoders. IEEE Transactions on Multimedia 21, 3136–3149.

[70] Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2242–2251.