DeepFacePencil: Realistic Face Image Creation from Free-hand Sketches

Anonymous Author(s)
Submission Id: 1570

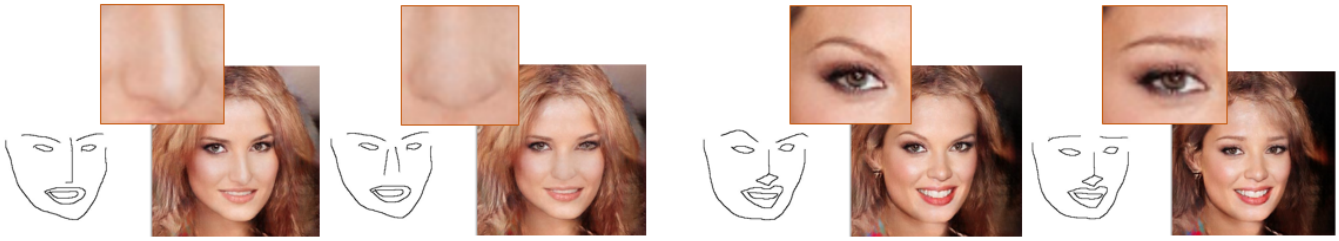


Figure 1: This is a teaser

ABSTRACT

In this paper, we explore the task of generating photo-realistic face images from sketches based on image-to-image translation framework. Since there exists no large-scale dataset of face sketches, existing methods utilize edge maps of face images as training data. However, edge maps perfectly align with edges of the corresponding face images, which limit existing models' generalization to hand-drawn sketches with vast stroke diversity. To address this problem, we propose a robust sketch-to-face translation model which is able to generate photo-realistic face images from hand-drawn sketches. A novel module, named spatial attention pooling (SAP) is designed to adaptively handle stroke diversity. We conduct extensive experiments on CelebA-HQ dataset. The experiment results show the superiority of our model over existing methods on perceptual realism and generalization.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**.

KEYWORDS

Image synthesis, spatial attention, sketch-based interface, face editing, conditional generative adversarial networks

ACM Reference Format:

Anonymous Author(s). 2020. DeepFacePencil: Realistic Face Image Creation from Free-hand Sketches. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '20), October 21–25, 2020, Seattle, US.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

MM '20, October 21–25, 2020, Seattle, US

ACM ISBN 978-1-4503-XXXX-X/20/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Flexibly creating new content is one of the most important goals in both computer graphics and computer-human interaction. While sketching is an efficient and natural way for common users to express their ideas for designing and editing new content, sketch-based interaction techniques have been extensively studied [**? ? ? ? ?**]. Imagery content is the most ubiquitous media with a large variety of display devices everywhere in our daily life. Creating new imagery content is one way to show people's creativity and communicate smart ideas. In this paper, we target portrait imagery, which is inextricably bound to our life, and present a sketch-based system, *DeepFacePencil*, which allows common users to create new face imagery by specifying the desired facial shapes via free-hand sketches.

Deep learning techniques have brought significant improvement on the realism of virtual images. Recently, a large amount of studies have been conducted on general image-to-image translation which aims to translate an image in one domain to a corresponding image in another domain, preserving the same content, such as structure, scene or objects [6, 9, 17, 19**?** –21]. Treating sketches as the source domain and realistic face images as the target domain, this task is a typical image-to-image translation problem. However, exiting image-to-image translation techniques are not off-the-shelf for this task due to the underlying challenges: data scarcity in the sketch domain and ambiguity in freehand sketches.

Since there exists no large-scale dataset of paired sketch and face images? and collecting hand-drawn sketches is time-consuming, existing methods [6, 12, 17] utilize edge maps or contours of real face images as training data when applied on the sketch-to-face task. Edge maps and contours enable existing models to be trained in a supervised manner and obtain plausible results on synthesized edge maps or contours . However, models trained on synthesized data are not able to achieve satisfactory results on hand-drawn sketches, specially on those drawn by common users without considerable drawing skills.

Since strokes in edge maps and contours align perfectly with edges of the corresponding real images, models trained on edge-aligned data tend to generate unreal shapes of facial parts following the inaccurate strokes when the input sketch is poorly drawn. Hence, for an imperfect hand-drawn sketch, it is a trade-off between *the realism* of the synthesized image and *the correspondence* between input sketch and the edges of the synthesized image. Models with rigorous edge-alignment fails to be generalized to sketches with imperfect strokes.

Moreover, we observe that the balance between the trade-off mentioned above varies from one position to another across the image. In a portrait sketch, some facial parts might be well-drawn while the others not. For the well-drawn facial parts, the balance are supposed to move towards the correspondence ensuring those parts in synthesized image depicting the user's imagination. On the other hand, the areas of poorly-drawn parts should emphasize the realism and not follow the irregular shapes and strokes.

«««< Updated upstream Based on the discussion above, we propose a novel sketch-based synthesis framework which is robust to hand-drawn sketches. A new module, named spatial attention pooling (SAP), is designed to adaptively adjust the spatially varying balance between *realism* and *correspondence* (()alignment?) across the image. In order to break the edge-alignment between sketches and real images, our SAP relaxes strokes with one-pixel widths to multiple-pixel widths using pooling operators. A larger width of a stroke, which is controlled by the kernel size of pooling operator, indicates the less restrict between this stroke and the corresponding edge in the synthesized image. However, the kernel size is not trainable using back propagation algorithm. Hence, for an input sketch, multiple branches of pooling operators with different kernel sizes are added in SAP to get multiple relaxed sketches with different widths. The relaxed sketches are then fused by a spatial attention layer which adjusts the balance of *realism* and *correspondence*. For different location in a portrait sketch, the spatial attention layer assigns high attention to the relaxed sketch with large width if this position requires more *realism* than *correspondence*. ======= Based on the discussion above, we propose a novel sketch-based synthesis framework which is robust to hand-drawn sketches. A new module, named spatial attention pooling (SAP), is designed to adaptively adjust the spacial-variant balance between *realism* and *correspondence* across the image. In order to break the edge-alignment between sketches and real images, SAP relaxes strokes with one-pixel widths to multiple-pixel widths using pooling operators. A stroke with a larger width, which is controlled by the kernel size of pooling operator, indicates the less restrict between this stroke and the corresponding edge in the synthesized image. However, the kernel size is not trainable using back propagation algorithm. Hence, for an input sketch, four branches of pooling operators with different kernel sizes are added in SAP to get four relaxed sketches with different widths. The relaxed sketches are then fused by a spatial attention layer which adjusts the balance of *realism* and *correspondence* . For each position, the spatial attention layer assigns high attention to the relaxed sketch with large width if this position requires more *realism* than *correspondence*. »»»» Stashed changes

In summary, our contribution in this paper is three-fold.

(a) Real Image (b) Edge Map (c) Contour (d) Semantic Boundary (e) Deformed Boundary

Figure 2: Comparison between a sketch generated from edge detection and from semantic boundary.

- Based on comprehensive analysis on the edge alignment issue in image translation frameworks, we propose a sketch-to-face translation system that is robust to hand-drawn sketches with various drawing skills.
- A novel deep neural network module for sketch, named *spatial attention pooling*, is designed to adaptively adjust the spatial-variant balance between the realism of the synthesized image and the correspondence between the input sketch and the synthesized image.
- Extensive experiments demonstrate the superiority of our model over existing methods on perceptual realism and generalization on the sketch-to-image task.

2 RELATED WORK

Our method is related to studies on image-to-image translation, sketch-based image generation and face image generation and editing. In this section, we discuss the most related works of our method.

2.1 Image-to-Image Translation

Given an input image from one domain, an image-to-image translation model outputs a corresponding image from another domain and preserves the content in the input image. Existing image-to-image translation models are based on generative adversarial networks conditioned on images. Pix2pix [6] is the first general image-to-image translation model which is able to be applied to different scenarios according to the paired training images, such as, semantic maps to real images, day images to night images, image coloring, and edge maps to real images. [7] utilizes semantic label maps and attributes of outdoor scenes as input and generates the corresponding photo-realistic images. In order to model multi-modal distribution of output images, BicycleGAN [**?**] encourages the connection between the output and the latent code to be invertible. CycleGAN [**?**], DualGAN [**?**], and DiscoGAN [**?**] propose unsupervised image translation model with a similar idea named cycle consistency borrowed from language translation literature. Pix2pixHD [17] is proposed as a high-resolution image-to-image translation model for generating photo-realistic image from semantic label maps using a coarse-to-fine generator and a multi-scale discriminator. It can also be applied to edge-to-photo generation by using the paired edge maps and photos as training data. However, the gap between edge maps and hand-drawn sketches challenges the generalization of these models.

2.2 Sketch-based Image generation

Sketch-based image generation is a hot topic in computer vision and computer graphics. Given a sketch of a scene with text labels

for objects, traditional methods, such as Sketch2Photo [2] and PhotoSketcher [3], search corresponding image patches from a large image dataset and then fuse the the retrieved image patches together according to the sketch. These methods are not able to ensure the global consistency of the resultant image and fails to generate totally new images. After the breakthrough made by deep neural networks (DNN) in computer graphics and computer vision, a variety of DNN-based models have been proposed for sketch-based image generation. The general image-to-image translation models mentioned above are able to be applied to sketch-based image generation once sketches and the corresponding images are used as training data. Besides, a few other models are designed specially for sketch inputs. SketchyGAN [**?**] aims to generate real images from multi-class sketches. A novel neural network module, called mask residual unit (MRU), is proposed to improve the information flow by injecting the input image at multiple scales. Edge maps are extracted from real images and utiled as training sketches. However, the resultant images of SketchyGAN are still not satisfied. Lines2face [12] utilizes a conditional self-attention module to preserve the completeness of global facial structure in generated face images. However, this model cannot be generalized to hand-drawn sketches directly.

2.3 Face Image Generation and Editing

Recently studies on face image generation and editing have made tremendous progress. The original generative adversarial network (GAN)'[4] is able to applied to generate face images from noise vectors. DCGAN [14] proposes a convolutional network to stabilize the training of GAN. PGGAN[8] utilizes a progressively growing architecture to generate high-quality face images of high resolution. Inspired by style transfer literature, StyleGAN [**?**] introduces a novel generator which synthesizes plausible high-resolution face images and learns unsupervised separation of high-level attributes and stochastic variation in synthesized images.

On the other side, a number of works focus on face image editing through different control information. StarGAN [**?**] designs a one-to-many translation framework which switches face attributes assigned by an attribute code. FaceShop [**?**] and SC-FEGAN [**?**] treats sketch-base face image editing as a sketch-guided image inpainting problem where stoke colors is also applied as guidance information.

3 DEEP NETWORK FOR SKETCH-PHOTO TRANSLATION

In this paper, we propose a sketch-to-photo translation model that is robust to hand-drawn sketches. In order to handle hand-drawn sketches, we design a novel spatial attention pooling (SAP) module to adaptively adjust the spatial-variant balance between *the realism* of the synthesized face image and *the correspondence* between input sketch and the edges in synthesized image. We arrange this section as follow. We first introduce the dataset we construct for our model in Subsection 3.1. Then we describe the architecture of our model in Subsection **??** and the proposed SAP in Subsection 3.3. At last we discuss losses applied in our model in Subsection 3.4 and the multi-stage training schedule in Subsection 3.5.

3.1 Face Sketches and Stroke Deformation

Paired face sketch-photo dataset is required for supervised sketch-to-face translation methods. Since there exits no large-scale paired sketch dataset, the training sketches used by existing methods [6, 12] are generated from face image dataset, e.g. CelebA-HQ face dataset, using edge detection algorithm such as HED [18]. However, the level of details in edge maps rely heavily on the value of a threshold of edge detection algorithm. An edge map with a large threshold contains too many redundant edges while an edge map with a small threshold fails to preserve the entire global facial structure [12].

Pix2pixHD [17] introduces another method to generate sketches from face images. Given a face image, the face landmarks are detected using an off-shelf landmark detection model. A new kind of sketch, denoted as *face contour*, is obtained by connect specific landmarks. However, since the pre-defined face landmarks mainly depict the facial area, a sketch-to-face model trained by face contours fails to generalize to hand-drawn sketches with hair, beard, or ornaments .

Based on the discussion above, we utilize a new kind of generated sketches with the assist of semantic maps. The CelebAMask-HQ dataset [11] provides a face semantic map for each face image in CelebA-HQ dataset. We basically use the boundary map of the semantic map as the sketch of the corresponding face image. Figure 2 shows an example of comparison between an edge map (b), a face contour (c) and a sketch generated from semantic boundary (d) from the same real image (a).

Stroke Deformation. A shortcut of sketches generated from semantic boundary (and those generated by edge detector) is that lines of sketches are perfectly aligned to edges of the corresponding face images. In order to break the edge-alignment between sketches and the corresponding real images and mimic the strokes of hand-drawn sketches, we apply a deformation to the lines, using a method similar to that in FaceShop [**?**]. Specifically, we vectorize lines of each sketches using AutoTrace algorithm [**?**]. Then offsets randomly selected from $[-d, d]^2$ are added to the control points and end points of the vectorized lines, where d is the maximum offset and we set $d = 11$ in our experiments unless specifically mentioned. We use the semantic boundary map as edge-aligned sketch, denoted as S, and semantic boundary map with random deformation as deformed sketch, S'.

3.2 Overview

The architecture of the proposed model is shown in Figure 3. Our model consists of two generators and one multi-scale discriminator. The first generator G generate realistic face image $G(S)$ from an edge-aligned sketch S while the second generator G' generate $G'(S')$ from a deformed sketch S'. Both generators are encoder-residual-decoder architectures, including a downsample encoder, several residual blocks and an upsample decoder, which is proven to be effective for generate high-quality images. The proposed spatial attention pooling module (SAP) is added before the encoder E' of G' to adaptively adjust the spatial-variant balance between *the realism* and *the correspondence*. The generators share weights of the residual blocks and decoders since the high-level features of both generators are supposed to share with each other.

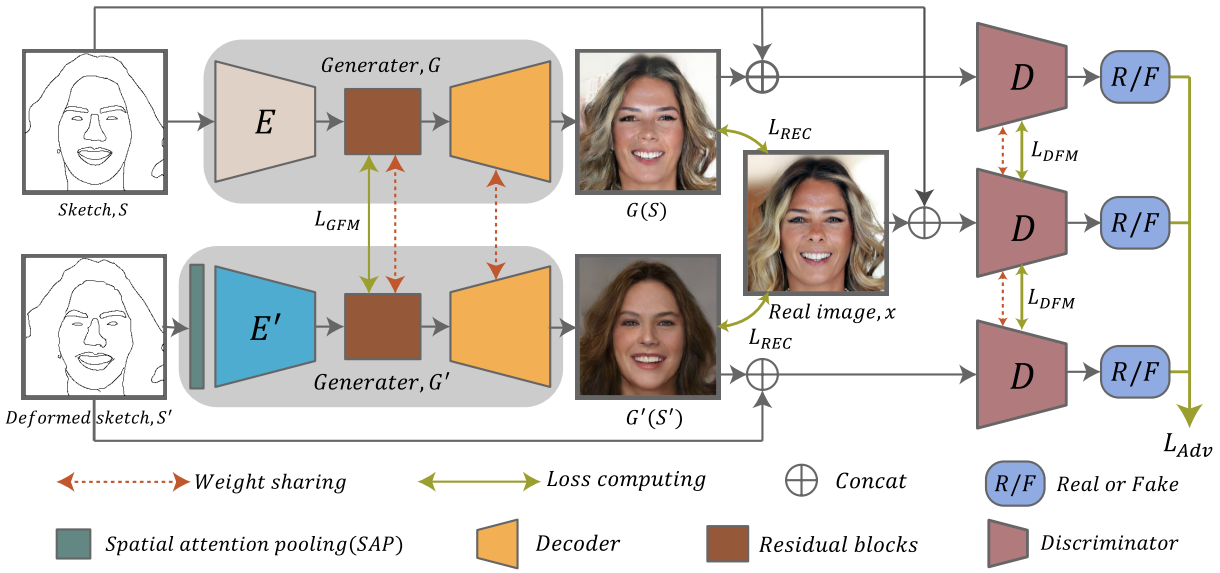


Figure 3: The architecture of our model.

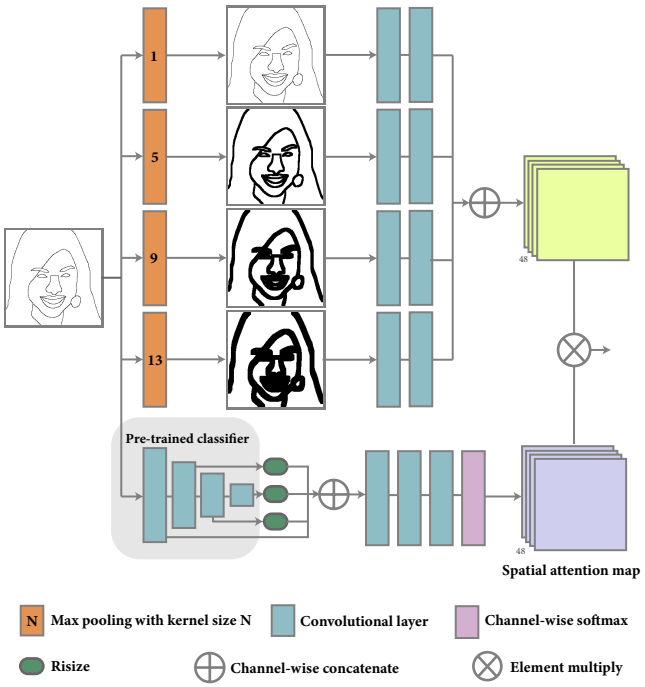


Figure 4: Sap

The discriminator is a multi-scale discriminator [17] which distinguishes real face images from generated fake images in both global and local scales. For the discriminator, the generated image $G(S)$ and $G'(S')$ concatenated with their input sketch S and S' respectively are treated as fake samples while a real face image sampled from real face distribution concatenated with its corresponding sketch is regarded as a real sample.

In order to guide the model with SAP to be tolerant with line distortion of deformed sketches, we design a novel generator feature matching loss for our task, besides the adversarial loss, the reconstruction loss and the discriminator feature matching loss used in pix2pixHD. The model is trained in a multi-stage training schedule to ensure the convergence of the training.

3.3 Spatial Attention Pooling

(Check if the idea about the relax is well-described) (Spatial-specific?Position-specific?Spatially variant?) A sketch-to-image model trained with edge-aligned sketch-image pairs tends to generate images whose edges align with the stokes in the input sketches. When an input hand-drawn sketch is not well-drawn, line distortions in the input sketch damages the quality of the generated face image. It is a trade-off between the realism of the output face image and the correspondence between the input sketch and the output face image.

In order to alleviate the edge alignment between the input sketch and the output face image, we desire to relax thin strokes to ambiguity bands with various width. One of the straightforward ways is to smooth the strokes with one-pixel width to multi-pixel width using image smoothing algorithm. However, the capacity of this hand-crafted way is limited, because the uniform smoothness for all positions of the whole sketch violate the unevenness of hand-drawn sketches on depicting different facial parts. We argue that the balance between the realism and the correspondence differs from one position to another across the face image. Therefore, the relax degree should be spatial-variant.

Based on the discussion above, we propose a new module, called spatial attention pooling (SAP), to adaptively relax the widths of strokes in the input sketch in a spatial-specific way. In order to break the edge-alignment between sketches and real images, SAP

relaxes strokes with one-pixel width to multiple-pixel width using pooling operators. A stroke with a larger width indicates the less restrict between this stroke and the corresponding edge in the synthesized image. The widths are controlled by the kernel sizes of pooling operators. However, the kernel size of pooling operator is not trainable using back propagation algorithm. SAP applies multiple branches of pooling operators with different kernel sizes to get multiple relaxed sketches with different widths. The relaxed sketches are then fused by a spatial attention layer which spatially adjusts the balance of *realism* and *correspondence*. The module is formulated as follow.

The architecture of SAP is shown in Figure 4. Given an input deformed sketch $S' \in \mathbb{R}^{H \times W}$, we first pass it through N_r pooling branches with different kernel sizes of $\{r_i, i = 1, \ldots, N_r\}$ to get $¶_i = Pooling_{r_i}(S')(i = 1, \ldots, N_r)$. Then we utilize convolutional layers to extract feature maps of P_i separately. These feature maps are concatenated to get a relaxed representation of S', denoted as R:

$$R = Cat(Conv_1(P_1), Conv_2(P_2), ..., Conv(P_{N_r})), \tag{1}$$

where $Conv_i()(i = 1, \ldots, N_r)$ indicates convolutional layers, Cat is a channel-wise concatenate operator.

On the other hand, we compute a spatial attention map A which controls the relax degrees of all positions by assigning different attention weights to R. A stoke with a large distortion is supposed to be assigned with a large relax degree. Hence, A is supposed to adaptively pay more attention (a large weight) to a $Conv_i(P_i)$ with a large kernel size in the areas with large line distortions. A straightforward way to get A is passing the input sketch through a few convolutional layers and these convolutional layers are trained to detect the areas with line distortions. However, we found the a few convolutional layers are insufficient to learn to detect line distortions directly. Therefore, we introduce a two-class classifier to ease the detection. Specifically, we pre-train a fully-convolutional two-class classifier C with three convolutional layers to distinguish sketches from deformed sketches. Then we utilize this pre-trained classifier to extract features of the input sketch S to get $C_i(S), i = 1, 2, 3$, where $C_i()$ denotes the ith feature maps extracted by C. These feature maps from classifier emphasize the differences between sketches and deform sketches. We resize and concatenate these feature maps, and pass them through three convolutional layers to get the spatial attention map:

$$A = Softmax(Conv([C_1, Up_2(C_2), Up_4(C_3)])), \tag{2}$$

where Up_2 and Up_4 indicates 2× and 4× upsampling, Conv(·) indicates three cascaded convoutional layers, and $Softmax()$ is a softmax layer computed over channels to ensuring that for each position of A, the sum of weights of all channels equals to 1.

At last, the output SAP is computed as:

$$SAP(S') = A * R, \tag{3}$$

where $*$ is element-wise multiplication.

3.4 Losses

generator feature matching effect and losses summary

(Before describe how you do it, please describe why you do this.)

Let $G_q(\cdot)$ produces the feature maps of the q-th layer in the generator G. Given an input sketch S and the corresponding deformed sketch $\tilde{S}$, we compute the generator feature matching loss as:

$$\mathcal{L}_{GFM}(G) = \mathbb{E}_{S \sim p_{data}(S)} \frac{1}{N_Q} \sum_{q \in Q} \frac{1}{|G_q|} \|G_q(S) - G_q(\tilde{S})\|_1, \tag{4}$$

where $|G_q|$ denotes the number of elements in $G_q(\cdot)$, Q indicates a set of the selected generator layers for computing this loss and the size of Q is N_Q. We select the xxx layers of the generator in our experiments.

Besides the generator feature matching loss $\mathcal{L}_{GFM}(G)$, for generator G and multi-scale discriminator $D = D_k | k = 1, 2, ..., N_D$, the adversarial loss $\mathcal{L}_{GAN}(G, D)$ and the discriminator feature matching loss $\mathcal{L}_{DFM}(G, D)$ are computed as the same form as those in pix2pixHD [17]. Discussion: add equations of these two losses or not The objective of the proposed model is:

$$\min_G \max_D \mathcal{L}_{GAN}(G, D) + \lambda \mathcal{L}_{DFM}(G, D) + \mu \mathcal{L}_{GFM}(G). \tag{5}$$

where λ and μ are the weights for balancing different losses. We set $\lambda = xxx$ and $\mu = xxx$ in our experiments.

3.5 Training Schedule

In order to train our model more stably, we introduce a multi-stage training schedule.

4 EXPERIMENTS

We propose a novel sketch-to-face translation model which is robust to hand-drawn sketches. We conduct extensive experiments to demonstrate the effectiveness of our model in generating high quality realistic face image from sketches.

Implementation Details. We implement our model on Pytorch [**?**]. Both generators share an encoder-residual-decoder structure except that an SAP is added to the front of the . The encoder contains four downsample convolutional layers, while the decoder contains four upsample convolutional layers. Nine residual blocks are added between the encoder and decoder to enlarge the capacity of generator. Weights of residual blocks and decoders of two generators share with each other. The multi-scale discriminator consists of three sub-networks for three scales separately. Each sub-network contains four downsample convolutional layers. Instance normalization [**?**] is applied after convolutional layers to stabilize training. ReLU [**?**] is used as activation for generators and LeakyReLU [**?**] for discriminator.

Training Details. All the networks are trained by Adam optimizer [10] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The initial learning rate is set to 0.0002 for each training stage and starts decay at the half of each stage. Batch size is 32. The entire training schedule takes about three days on four NVIDIA GTX 1080Ti GPUs with 11GB GPU memory.

Data. CelebA-HQ [8] is a large-scale face image dataset which contains 30K 1024×1024 high-resolution face images. We use face images in this dataset as real images. CelebAMask-HQ [11] offers manually-annotated face semantic masks for CelebA-HQ with

19 classes including all facial components and accessories such as skin, nose, eyes, eyebrows, ears, mouth, lip, hair, hat, eyeglass, earring, necklace, neck, and cloth. We utilize semantic masks in this dataset to extract semantic boundary maps as edge-aligned sketches. Both real images and sketches are resized to 256×256 in our experiments.

Baseline Model. Pix2pixHD [17] is a state-of-the-art image-to-image translation model for high-resolution images. With the edge-aligned sketches and face real images, we train pix2pixHD with its low-resolution version of generator ('global generator') as a baseline model in our experiment, denoted as *baseline*. In order to conduct a fair comparison on generalization, we also train the baseline model with both edge-aligned sketches and deformed sketches, denoted as *baseline_deform*.

4.1 Evaluation Metrics

Evaluating the performance of generative models has been studied in image generation literature. It is proven to be a complicated task because a model with good performance with respect to one criterion does not necessarily imply good performance with respect to another criterion [13]. A proper evaluation metrics should be able to present the joint statistics between conditional input samples and generated images which means traditional metrics, such as pixel-wise mean-squared error, do not reveal the performance of generative models. Therefore, we utilize three popular quantitative perceptual evaluation metrics based on deep neural features: Inception Score (IS) [15], Fréchet Inception Distance (FID) [5], and Kernel Inception Distance (KID) [1]. These metrics are proven to be consistent with human evaluation in assessing the realism of images.

Inception Score (IS). IS applies an Inception model pre-trained on ImageNet to extract features of generated images and computes the KL divergence between the conditional class distribution and the marginal class distribution. We note that IS is reported to biased in some cases because its evaluation is based more on the recognizability rather than on the realism of the generated samples [16]. Higher IS presents higher quality of generative images.

Fréchet Inception Distance (FID). FID is a recently proposed evaluation metrics for generative models and proven to be consistent with human evaluation in assessing the realism of images. FID computes Wasserstein-2 distance between features of generated images and real images which are also extracted by a pre-trained Inception model. Lower FID indicates that the generative distribution is closer to the real distribution.

Kernel Inception Distance (KID). KID measures the distance of two distributions by calculating the squared maximum mean discrepancy between Inception features. KID is pointed out to an unbiased estimator with a cubic kernel [1]. Models with better performance are supposed to achieve a lower KID.

4.2 Generative Quality Comparison with Image Translation networks

Existing image-to-image translation models can be trained for sketch-to-face translation using the paired dataset. Since the quality of the generated images presents the basic performance of a generative model, we first compare the generative quality between our model and existing models using three evaluation metrics of IS, FID, and KID. In this experiment, existing models are trained by edge-aligned face sketches and the corresponding real face images. We test both our model and existing models with edge-aligned sketches in the test set. Two more existing methods are included in this experiments besides the baseline model. Pix2pix [6] is the first general image-to-image translation framework which is able to applied to a variety of applications by switching the training data. We use the default setting to train pix2pix model with paired edge-aligned sketches and real face images. Lines2face [12] is proposed as a sketch-to-face translation model which aims to preserve the entire facial structure by introducing self-attention mechanism to image-to-image translation framework. Lines2face is original trained with edge maps. We only switch the training data to paired edge-aligned sketches and real face images, leaving other settings unchanged.

Table 1 shows the quantitative results of this experiment. The our model surpasses all existing models with respect to all three evaluation metrics and xxxxxxxxxx.

Visual results are shown in Figure 5

Table 1: Results of generative quality comparison.

	Pix2pix [6]	Lines2face [12]	Baseline [17]	Baseline_deform	Ours
IS	–	–	–	–	–
FID	–	–	635.87	625.98	**606.7**
KID	–	–	–	–	–

4.3 Generalization Comparison with Baseline Model

In order to verify the generalization ability of our model, we design several experiments and compare our model with the baseline model by testing with sketches of different levels of deformation, well-drawn sketches and poor-drawn sketches.

4.3.1 Different Levels of Deformation. As mentioned in Subsection 3.1, we deform a edge-aligned sketch S to obtain a corresponding deformed sketch S' by vectorizing strokes in S and adding random offsets to the control points and end points of the vectorized strokes. The maximum offset d is set to 11 in the training data. We further create more deformed sketches with different levels of deformation, denoted as S'_d, by modifying the maximum offset d, where d indicates the level of deformation. We examine the generalization ability of our model and baseline model on these sketches. Note that the *baseline* is trained with only edge-aligned sketches while our model and *baseline_deform* model are trained both edge-aligned sketches and deformed sketches with $d = 11$.

In this experiment, the input sketches are deformed by large offsets where the maximum d is set to 30. As shown in Figure 6, strokes in the largely deformed sketches are quite different from those in the training sketches including edge-aligned sketches S and deformed sketches $S_1 1'$. xxxxx.

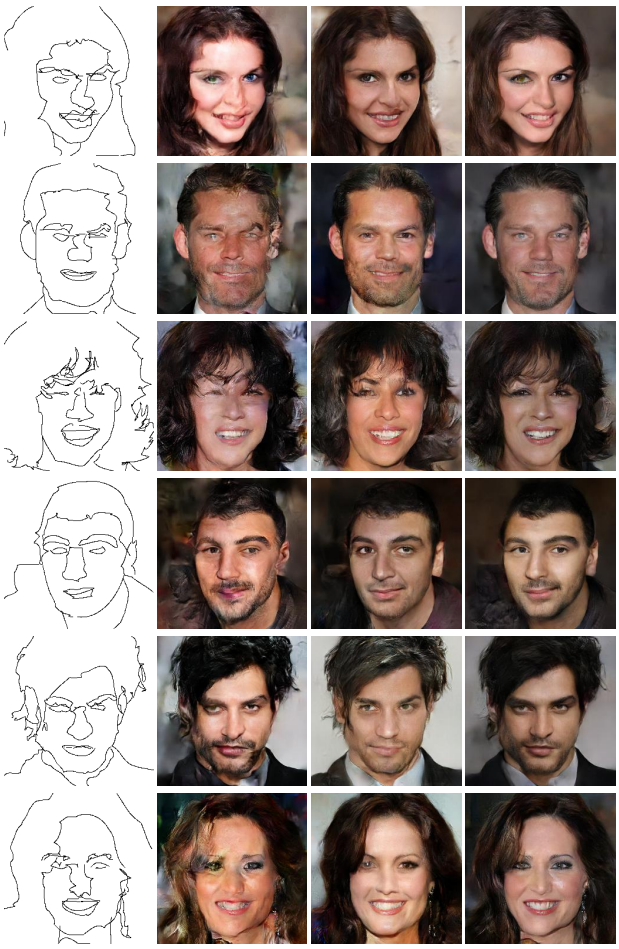
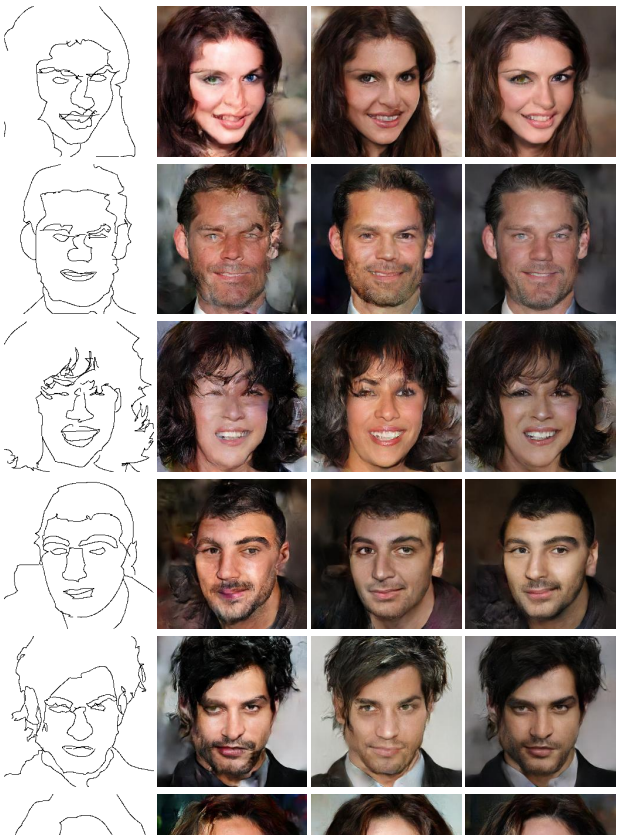


Figure 6: Result analyze



4.3.2 Well-Drawn Sketches. We invite professional users with well-trained drawing skills to draw sketches for testing. These professional users are asked to draw sketches on the drawing board to mimic the training sketches. We compare our model with baseline model by testing with

4.4 Limitations and future work

REFERENCES

[1] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. In *6th International Conference on Learning Representations, ICLR 2018.* https://openreview.net/forum?id=r1lUOzWCW

[2] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. 2009. Sketch2Photo: internet image montage. *ACM Trans. Graph.* 28, 5 (2009), 124:1–124:10.

[3] M. Eitz, R. Richter, K. Hildebrand, T. Boubekeur, and M. Alexa. 2011. Photosketcher: Interactive Sketch-Based Image Synthesis. *IEEE Computer Graphics and Applications* 31, 6 (Nov 2011), 56–66.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada.* 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems 30.* 6626–6637.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 5967–5976. https://doi.org/10.1109/CVPR.2017.632

[7] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. 2016. Learning to Generate Images of Outdoor Scenes from Attributes and Semantic Layouts. *CoRR* abs/1612.00215 (2016). arXiv:1612.00215

[8] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *6th International Conference on Learning Representations, ICLR 2018.*

[9] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192* (2017).

[10] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015,*. http://arxiv.org/abs/1412.6980

[11] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[12] Yuhang Li, Xuejin Chen, Feng Wu, and Zheng-Jun Zha. 2019. LinesToFacePhoto: Face Photo Generation From Lines With Conditional Self-Attention Generative Adversarial Networks. In *Proceedings of the 27th ACM International Conference*

on Multimedia (Nice, France) *(MM '19)*. ACM, New York, NY, USA, 2323–2331. https://doi.org/10.1145/3343031.3350854

[13] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. [n.d.]. Are GANs Created Equal? A Large-Scale Study. In *Advances in Neural Information Processing Systems 31*. 700–709.

[14] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *4th International Conference on Learning Representations, ICLR 2016,*. http://arxiv.org/abs/1511.06434

[15] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. 2016. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems 29*. 2234–2242.

[16] Lucas Theis, Aaron van den Oord, and Matthias Bethge. 2016. A note on the evaluation of generative models. In *4th International Conference on Learning Representations, ICLR 2016,*.

[17] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8798–8807. https://doi.org/10.1109/CVPR.2018.00917

[18] Saining Xie and Zhuowen Tu. 2015. Holistically-Nested Edge Detection. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 1395–1403. https://doi.org/10.1109/ICCV.2015.164

[19] Z. Yi, H. Zhang, P. Tan, and M. Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2868–2876. https://doi.org/10.1109/ICCV.2017.310

[20] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2242–2251. https://doi.org/10.1109/ICCV.2017.244

[21] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017. Toward Multimodal Image-to-Image Translation. In *Advances in Neural Information Processing Systems 30*. 465–476.