DeepFacePencil: Creating Face Images from Freehand Sketches

Anonymous Author(s)
Submission Id: 1570

ABSTRACT

In this paper, we explore the task of generating photo-realistic face images from hand-drawn sketches. Existing image-to-image translation methods require a large-scale dataset of paired sketches and images for supervision. They typically utilize synthesized edge maps of face images as training data. However, these synthesized edge maps strictly align with the edges of the corresponding face images, which limit their generalization ability to real hand-drawn sketches with vast stroke diversity. To address this problem, we propose DeepFacePencil, an effective tool that is able to generate photo-realistic face images from hand-drawn sketches, based on a novel dual generator image translation network during training. A novel spatial attention pooling (SAP) is designed to adaptively handle stroke distortions which are spatially varying to support various stroke styles and different level of details. We conduct extensive experiments and the results demonstrate the superiority of our model over existing methods on both image quality and model generalization to handdrawn sketches.

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: Creating Face Images from Freehand Sketches. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '20), October 21–25, 2020, Seattle, US.* ACM, New York, NY, USA, 9 pages. 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 [2, 4, 9, 24, 31]. 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 improvements 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 [10, 15, 27, 30, 32, 33]. 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 [10, 18, 27] 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 poor-drawn. Hence, for an imperfect hand-drawn sketch, it is a trade-off between *the realism* of the synthesized image and *the conformance* between input sketch and the edges of the synthesized image. Models with high 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 conformance 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.

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 *conformance* 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

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

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

Figure 1: Comparison between a sketch generated from edge detection and from semantic boundary.

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 *conformance*. 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 *conformance*.

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

- 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 spatially varying balance between the realism of the synthesized image and the conformance 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 [10] is the first general image-to-image translation model which can be applied to different scenarios according to paired training images, such as, semantic maps to natural images, day images to night images, image colorization, and edge maps to images. [12] 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 [33] encourages the connection between the output and the latent code to be invertible. CycleGAN [32], DualGAN [30], and DiscoGAN [15] propose unsupervised image translation models with a common idea named cycle consistency, which is borrowed from language translation literature. Pix2pixHD [27] 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 when trained on paired edge maps and photos. However, the large gap between synthesized 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 multimedia and computer vision. Given a sketch describe the desired scene layout with text labels for objects, traditional methods, such as Sketch2Photo [1] and PhotoSketcher [6], search image patches from a large-scale image dataset and fuse 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. Nevertheless, it is challenging for these methods to ensure global consistency of the resultant images. Thus they frequently fail to generate totally new images. After the breakthrough made by deep neural networks (DNNs) in many image understanding tasks, a variety of DNN-based models have been proposed for sketch-based image generation. The general image-to-image translation models mentioned above can be easily extended to sketch-based image generation once sketches and their corresponding images are available as training data. Besides, a few other models are designed specially for sketch inputs. SketchyGAN [3] 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 utilized as training sketches. However, the resultant images of SketchyGAN are still not satisfied. LinesToFacePhoto [18] employs 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 due to distinct stroke characteristics.

2.3 Face Image Generation and Editing

Recently studies on face image generation and editing have made tremendous progress. Using generative adversarial network (GAN) [7], realistic face images can be generated from noise vectors. DCGAN [22] introduces a novel network to stabilize training of GAN. PGGAN[13] utilizes a progressively growing architecture to generate high resolution face images. Inspired by style transfer literature, StyleGAN [14] 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 [5] designs a one-to-many translation framework which switches face attributes assigned by an attribute code. FaceShop [21] and SC-FEGAN [11] treats sketch-base face image editing as a sketch-guided image inpainting problem where stoke colors is also applied as guidance information.

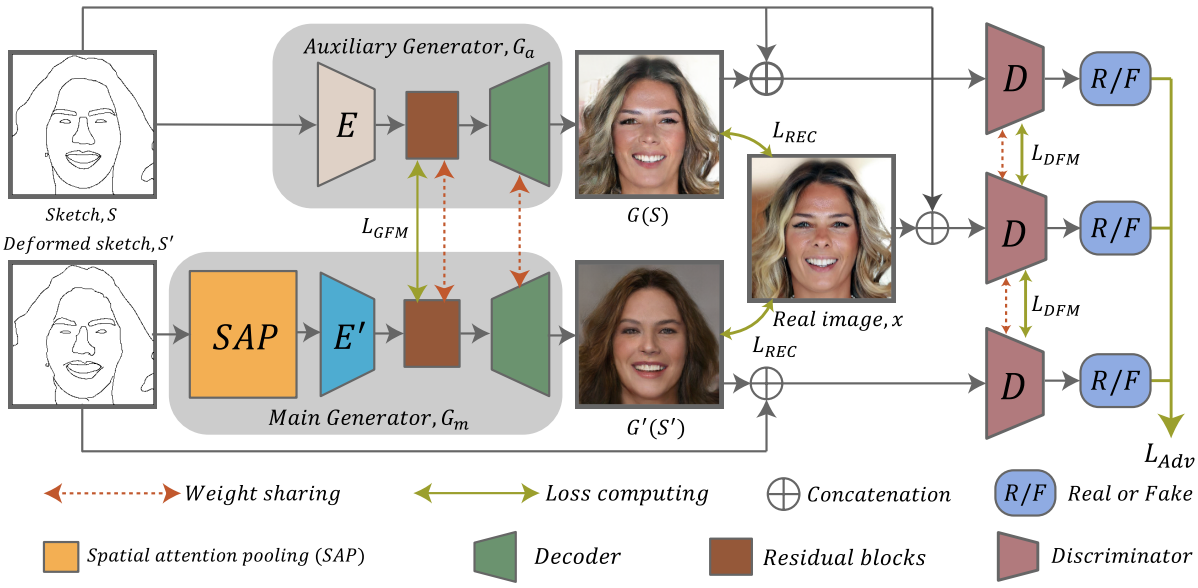


Figure 2: The architecture of our dual-generator model. In order to train a face image generator G_m for hand-drawn sketches, we synthesize deformed sketches S' from an edge-aligned sketch S and design a spatially attention pooling module to extract shape and structure features from distorted sketches. The dual generators G_m and G_m are trained simultaneously in an adversarial manner.

3 DEEP NETWORK FOR SKETCH-PHOTO TRANSLATION

The task of sketch-to-photo translation can be defined as looking for a generator $G(S)$ so that the generated image $x = G(S)$ from a hand-drawn sketch S looks realistic and keeps the shape characteristics for the input sketch. Existing image translation techniques train a neural network as the generator with paired of sketch and photo data $(\mathcal{S}, \mathcal{X})$. Due to the scarcity of real hand-drawn sketches, existing techniques synthesize sketches in a certain style to approximate the sketch set $\mathcal{S}$ from face image set $\mathcal{X}$ to train their generator in an adversarial manner. The synthesized sketches $\mathcal{S}_{syn}$ are usually well aligned with the face images and present different distributions from hand-drawn sketches $\mathcal{S}$. These models typically fail to generalize to hand-drawn sketches by common users.

As shown in Figure 2, G_m is the main generator trained by deformed sketches S', aiming to generate plausible photo-realistic face images from unseen hand-drawn sketches in test stage. G_a is an auxiliary generator trained with edge-aligned sketches whose goal is to guide G_m to adaptively sense the line distortion in deformed sketches.

We propose a novel network architecture with a specially designed training strategy to improve the capability of the sketch-based image generator. Figure 2 shows the overview of our method. In order to synthesize a set of sketches $\mathcal{S}_{syn}$ that has similar distribution with hand-drawn sketches $\mathcal{S}$, we deform the edge-aligned sketches to generate a set of deformed sketches $\mathcal{S}_{dfm}$ to augment the training set. We propose a novel framework using dual generators from the edge-aligned sketches $\mathcal{S}_{syn}$ and the deformed sketches $\mathcal{S}_{dfm}$ respectively. G_m is the main generator trained by deformed sketches S', aiming to generate plausible photo-realistic face images from unseen hand-drawn sketches in test stage. G_a is an auxiliary generator trained with edge-aligned sketches whose goal is to guide G_m to adaptively sense the line distortion in deformed sketches. A spatial attention pooling module (SAP) is added before the encoder E_m of G_m to adjust the spatially varying balance between *the realism* of generated images and *the conformance* between the generated image and the input sketch.

The dual generators are trained together under a set of supervision. First, for a triplet (S, S', x), both the generators G_m and G_a are trained to produce images $G_m(S')$ and $G_a(S)$ to approximate the real image x under a reconstruction loss L_{rec}. Second, a multi-scale discriminators D is employed for three combinations of sketch-image pairs to distinguish real face images from generated fake images in both global and local scales. By training the dual generators simultaneously with the discriminator adversarially, our main generator G_m effectively captures the spatially-varying stroke distortions and maps it to the manifold of well-drawn sketches to produce realistic face images.

3.1 Synthetic 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 [10, 18] are generated from face image dataset, e.g. CelebA-HQ face dataset, using edge detection algorithm such as HED [29]. 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 [18].

Pix2pixHD [27] 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 [17] 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 1 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 [21]. Specifically, we vectorize lines of each sketches using AutoTrace algorithm [28]. 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 S, and semantic boundary map with random deformation as deformed sketch S'.

3.2 Spatial Attention Pooling

A sketch-to-image model trained with edge-aligned sketch-image pairs tends to generate images whose edges strictly align with the stokes of the input sketch. 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 generated face image and the conformance 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 propose to relax thin strokes to a tolerance region with various width. A straightforward way is to smooth the strokes to multi-pixel width by image smoothing or dilation. 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 conformance differs from one position to another across the face image. Therefore, the relaxation degree should be spatially varying.

Based on the discussion above, we propose a new module, called spatial attention pooling (SAP), to adaptively relax the strokes in the input sketch to spatially varying tolerance regions. 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,

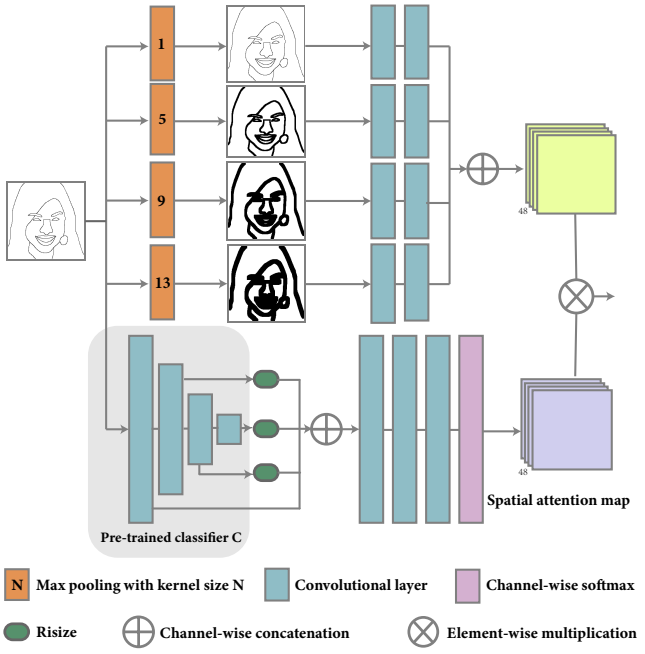


Figure 3: Network architecture of our spatial attention pooling module.

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 *conformance*. The module is formulated as follow.

The architecture of SAP is shown in Figure 3. 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:

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

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.

3.3 Losses

Our model consists of two generators, G_a for edge-aligned sketch S and G_m deform sketch S', and one discriminator D. Loss functions and objective of our model are discussed as follow.

Reconstruction Loss. For either generator, a reconstruction loss is applied to guide the generated image to get close to its corresponding real image x.

$$\begin{aligned}\mathcal{L}_{Rec}(G_a, G_m) = &\mathbb{E}_{(S,x)\sim p_{data}(S,x)}\|G_a(S)-x\|_1 \\ &+ \mathbb{E}_{(S',x)\sim p_{data}(S',x)}\|G_m(S')-x\|_1,\end{aligned} \tag{4}$$

Adversarial Loss. The multi-scale discriminator [27] D consists of three sub-discriminators $D_i, i = 1, 2, 3$. The adversarial loss for G and D is defined as:

$$\begin{aligned}\mathcal{L}_{adv}(G_a; D) = &\frac{1}{3}\sum_{i=1}^{3}\mathbb{E}_{(\boldsymbol{S},\boldsymbol{x})\sim p_{data}(\boldsymbol{S},\boldsymbol{x})}\left[\log D_i(\boldsymbol{S},\boldsymbol{x})\right] \\ &+ E_{\boldsymbol{x}\sim p_{data}(\boldsymbol{S})}\left[\log\left(1 - D_i\left(\boldsymbol{S}, G_a(\boldsymbol{S})\right)\right)\right].\end{aligned} \tag{5}$$

The adversarial loss for G_m and D, denoted as $\mathcal{L}_{adv}(G_m; D)$ is defined similarly.

Discriminator Feature Matching Loss. Similar to pix2pixHD[27] and lines2face [18], we use a discriminator feature matching loss as the perceptual loss, which is designed to minimize the error between generated image and real image in feature space. Here discriminator feature matching loss use the discriminator as the feature extractor. Let $D_i^q()$ be the output of qth layer in D_i. This loss is defined as:

$$\mathcal{L}_{DFM}(G_a) = \frac{1}{3N_Q}\mathbb{E}_{(\boldsymbol{S},\boldsymbol{x})\sim p_{data}(\boldsymbol{S},\boldsymbol{x})}\sum_{i=1}^{3}\sum_{q\in Q}\frac{1}{n_i^q}\|D_i^q(G_a(\boldsymbol{S}))-D_i^q(\boldsymbol{x})\|_1, \tag{6}$$

where Q is the selected layers of discriminator for computing this loss, N_Q is the number of elements in Q, n_i^q is the number of elements in D_i^q. Also, the discriminator feature matching loss for G_m and D, denoted as $\mathcal{L}_{DFM}(G_m)$, is defined similarly.

Generator Feature Matching Loss. Similar to discriminator feature matching loss which is designed to minimize the error between generated image and real image in feature space, we propose a novel generator feature matching loss that aims to minimize the error between the presentations of S and S' in generator feature space. Let $G^t()$ and $G_m^t()$ be the output of tth layer in G and G_m respectively. This loss is calculated as:

$$\mathcal{L}_{GFM}(G_a, G_m) = \mathbb{E}_{(S,S')\sim p_{data}(S,S')}\frac{1}{N_T}\sum_{t\in T}\frac{1}{|n^t|}\|G_a^t(S)-G_m^t(S')\|_1, \tag{7}$$

where T is the set of selected generator layers for calculating this loss, N_T is the number of elements of T, and n^t is the number of elements of $G_a^t()$ and $G_m^t()$. We select the output of the first four residual blocks of the generator in our experiments.

The objective of the proposed model is:

$$\min_G \max_D \mathcal{L}_{GAN}(G_a, D)+\mathcal{L}_{GAN}(G_m, D)+\lambda(\mathcal{L}_{DFM}(G_a, D)+\mathcal{L}_{DFM}(G_m, D))+\mu\mathcal{L}_{(} \tag{8}$$

where λ and μ are the weights for balancing different losses. We set $\lambda = 10$ and $\mu = 10$ in our experiments.

3.4 Training Schedule

In order to train our model more stably, we introduce a multi-stage training schedule. At the first stage, we use edge-aligned sketches and real images to train G_a and D without loss function related to G_m. At the second stage, we train SAP and the encoder of G_m and D from scratch while fixing weights of other parts. Note that the residual blocks and the decoder of G_m share weights with those of G_a and are kept unchanged in this stage. At the last stage, we finetune the whole model.

4 EXPERIMENTS AND DISCUSSIONS

Our method 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 drawn by different users with diverse painting skills.

4.1 Implementation Settings

Before showing experimental results, we first introduce details in our network implementation and training.

Implementation Details. We implement our model on Pytorch [20]. Both generators for edge-aligned sketches and deformed sketches share an encoder-residual-decoder structure with shared weights except that an SAP module is added to the front of the main generator G_m for deformed sketches. The encoder consists of four convolutional layers with 2× downsampling, while the decoder consist of four convolutional layers with 2× upsampling. Nine residual blocks between the encoder and decoder enlarge the capacity of the generators. The multi-scale discriminator D consists of three sub-networks for three scales separately, same as Pix2PixHD [27]. Instance normalization [26] is applied after the convolutional layers to stabilize training. ReLU is used as activation for generators and LeakyReLU for discriminator.

Data. To produce triplets, consisted of sketches, deform sketches, and real images, (S, S', x) for our network training, we use CelebA-HQ [13], a large-scale face image dataset which contains 30K 1024× 1024 high-resolution face images. CelebAMask-HQ [17] 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. Deformed sketches are generated by vectorizing and adding random offsets to edge-aligned sketches as discussed in the last section. Real images, sketches, and deformed sketches are resized to 256×256 in our experiments.

Training Details. All the networks are trained by Adam optimizer [16] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. For each training stage, the initial learning rate is set to 0.0002 and starts to decay at the half of training procedure. We set batch size as 32. The entire training process takes about three days on four NVIDIA GTX 1080Ti GPUs with 11GB GPU memory.

Baseline Model. Pix2pixHD [27] is a state-of-the-art image-to-image translation model for high-resolution images. With the edge-aligned sketches and real face 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 augmented dataset by adding pairs of deformed sketches and images, denoted as *baseline_deform*. The key idea of our method is using SAP and dual generators to improve the tolerance to sketch distortions. The local enhancer part of pix2pixHD, which is designed for high-resolution image synthesis, can be easily added to improve fine textures for both baseline models and our model in the future.

4.2 Quantitative Evaluation on Image Quality

4.2.1 Evaluation Metrics. Evaluating the performance of generative models has been studied for a period of time 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 [19]. A proper evaluation metrics should be able to present the joint statistics between conditional input samples and generated images. Traditional metrics, such as pixel-wise mean-squared error, can not effectively measure the performance of generative models. We utilize two popular quantitative perceptual evaluation metrics based on image features extracted by DNNs: Inception Score (IS) [23] and Fréchet Inception Distance (FID) [8]. These metrics are proven to be consistent with human perception 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. Higher IS presents higher quality of generative images. Note that IS is reported to be biased in some cases because its evaluation is based more on the recognizability rather than on the realism of generated samples [25].

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

4.2.2 Image Quality Comparison. Existing image-to-image translation models can be trained for sketch-to-face translation using paired sketch and image data. Since the quality of generated images presents the basic performance of a generative model, we first compare the quality of generated images by different generative models using IS and FID. We test these models with edge-aligned sketches that are synthesized from images in the test set. Besides the baseline model, we also test Pix2pix [10], which is the first general image-to-image translation framework. It can be 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. All these methods produce face image in dimension of 256×256.

Table 1 shows the quantitative evaluation results of four models. Our model surpasses other models by a small margin with respect to two evaluation metrics. Visual results are shown in Figure 4. As we can see, all of the four models are able to generate plausible face images from sketches of test dataset. Since these test sketches are generated using the same method as training sketches, the test data distribution is quite close to/overlapped with the training data distribution. Both our model and existing models are easily to be generalize to handle these samples. We will show our model's superiority on generalization to more challenging sketches in the next few experiments.

Table 1: Results of generative quality comparison.

	Pix2pix [10]	Baseline [27]	Baseline_deform	Ours
IS	2.186	2.298	2.369	**2.411**
FID	289.3	259.1	244.3	**242.1**

4.3 Comparison of Generalization Capability

In order to verify the generalization ability of our model, we compare our model with state-of-the-art image translation models by testing with synthesized sketches of different levels of deformation, well-drawn sketches and poorly-drawn sketches by novel users.

4.3.1 Different Level of Deformation. As mentioned in Subsection 3.1, we deform an edge-aligned sketch S to obtain a corresponding deformed sketch S' by adding random offsets to the control points and end points of the vectorized strokes in S. The maximum offset d is set to 11 in the training data. We further create more deformed sketches with multiple levels of deformation, denoted as S'_d, by modifying the maximum offset d. We examine the generalization ability of our model and baseline model on these deformed 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 $d = 11$.

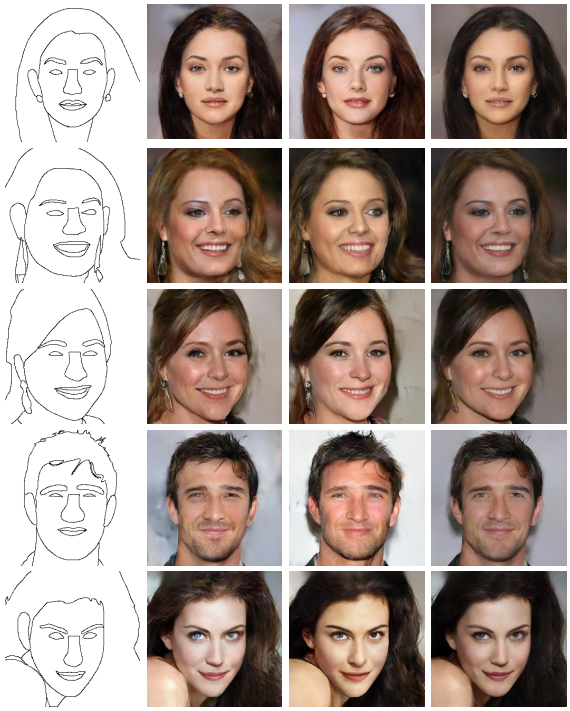


Figure 4: Both our model and existing models, which generate plausible photo-realistic face images from sketches in test set, are able to be generalized to test sketches from similar distribution as the training distribution. We will show our model's superiority on generalization to more challenging sketches in the next few experiments.

In this experiment, the input sketches are deformed by larger offsets where the maximum d is set to 30. As shown in Figure 5, strokes in sketches with larger deformation looks quite different from those in the training sketches including edge-aligned sketches S and deformed sketches S'_{11}. By adding deformed sketches into training data, *baseline_deform* produces better images than *baseline*. However, when larger deformation occurs, *baseline_deform* suffers from artifacts in facial features, for example, the mouth in the first example, the eyes of the third and the last case in Figure 5, In comparison, our model produces more realistic face images with more symmetric eyes and find textures, benefited from its ability of capturing shape feature from deformed strokes by our spatial attention pooling module.

4.3.2 Hand-Drawn Sketches. Besides the synthesized sketches with stroke deformation, we further examine the model generalization ability by comparing performances of our model with baseline model tested with two kinds of hand-drawn sketches: expert-drawn sketches and common-user sketches drawn by users without professional painting skills.

Expert-Drawn Sketches. We invite an expert with well-trained drawing skills to draw portrait sketches for testing. These expert sketches were drawn on a pen tablet so that the strokes are smooth and precise. Note that shading strokes are not drawn. Figure 6 shows a group of face images generated by different models from several expert-drawn sketches. Even with well-drawn strokes, the *baseline* and *baseline_deform* frequently fail to produce realistic textures and complete structures of eyes or mouths. In comparison, our results are more realistic with fine textures and intact structures.

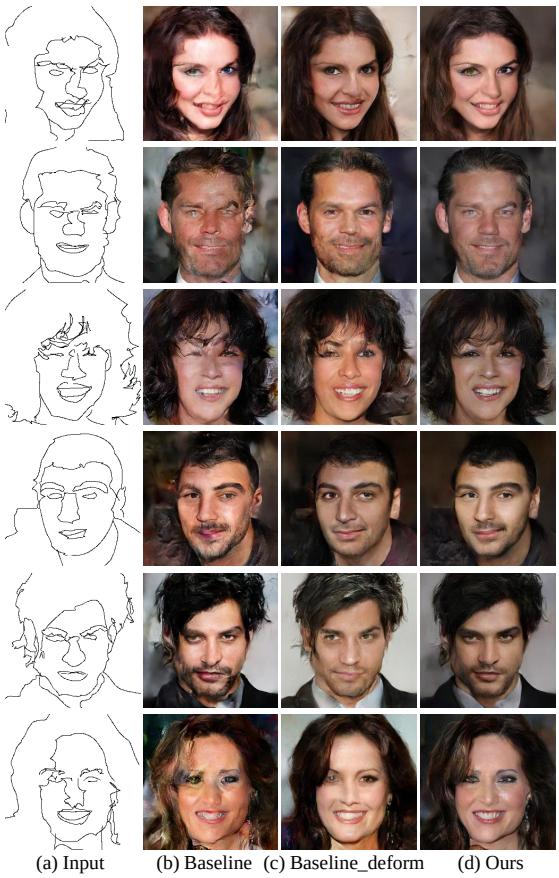
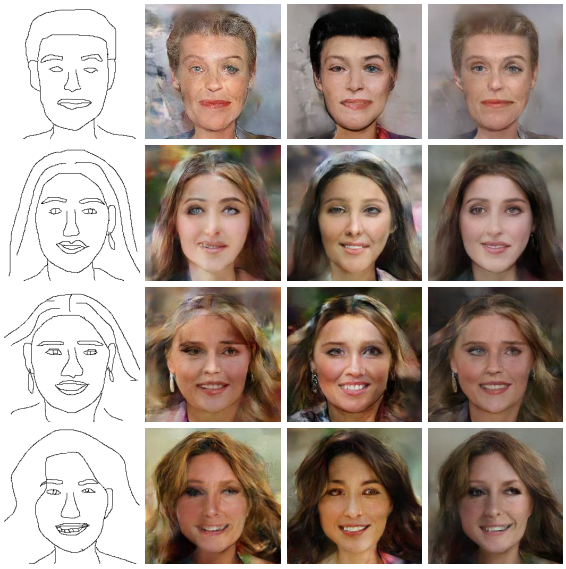


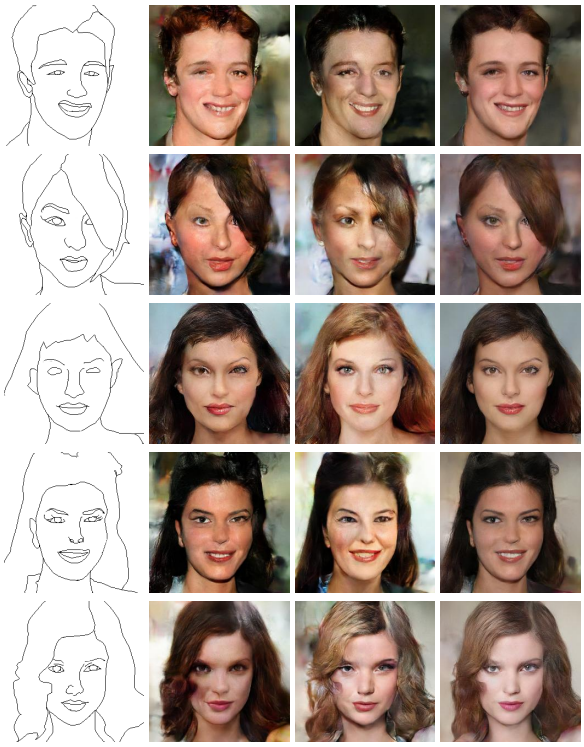
Figure 5: For sketches with large deformation, both *baseline* model and *baseline_deform* model fail to generate satisfying results, in which artifacts can be found in areas with large sketch deformation. Our results maintain high quality even large deformation occurs.

Common Sketches. We also invited a large number of graduate students without drawing skills to draw freehand sketches of their imagined faces using mouses. Hence, strokes of these common sketches rough depicts the desired face structure and shapes of facial features with some distortion. Moreover, common sketches turn to be of different levels of details. For example, some sketches contains many strokes inside the hair areas which are blank in the training sketches. Results shown in Figure 7 demonstrate that our model is robust to these poorly-drawn sketches. In constrast, the diversity of stroke styles and detail levels significantly damage the quality of the results from the *baseline* model and *baseline_deform* model.



(a) Input (b) Baseline (c) Baseline_deform (d) Ours

Figure 7: For these challenging sketches drawn by common users, our model still generate plausible results. Results of baseline models are over-blurred and contains obvious artifacts.



(a) Input (b) Baseline (c) Baseline_deform (d) Ours

Figure 6: Our model is able to be generalized to well-drawn expert sketches while results of baseline models are degenerated.

5 CONCLUSION

In this paper, we present DeepFacePencil, a novel deep neural network which allows common users to create photorealistic by freehand sketching. The robustness of our sketch-based face generator comes from the proposed dual generators and the spatial attention pooling module. The proposed spatial attention pooling module adaptively adjusts the spatially varying balance between the image realism and the conformance between the input sketch and the synthesized image. By adding the SAP module to our dual-generator network and training the two generators simultaneously to enforce the main generator to effectively capture face structure and facial feature shapes from coarsely drawn sketches. Extensive experiments demonstrate that our DeepFacePencil successfully produce high quality face images from freehand sketches drawn by users in diverse drawing skills.

REFERENCES

[1] 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.

[2] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. 2009. Sketch2Photo: Internet image montage. *ACM Trans. Graph* 27, 2 (2009), 124:1–124:10.

[3] Wengling Chen and James Hays. 2018. SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 9416–9425. https://doi.org/10.1109/CVPR.2018.00981

[4] Xuejin Chen, Sing Bing Kang, Ying-Qing Xu, and Julie Dorsey. 2008. Sketching reality: Realistic interpretation of architectural designs. *ACM Trans. Graph.* 27, 2 (2008), 11:1–11:15.

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *CVPR*. 8789–8797.

[6] 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.

[7] 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

[8] 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.

[9] Takeo Igarashi, Satoshi Matsuoka, and Hidehiko Tanaka. 1999. Teddy: A sketching interface for 3-D freeform design. In *Proc. SIGGRAPH*. 409–416.

[10] 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

[11] Youngjoo Jo and Jongyoul Park. 2019. SC-FEGAN: Face Editing Generative Adversarial Network With User's Sketch and Color. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 1745–1753. https://doi.org/10.1109/ICCV.2019.00183

[12] 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

[13] 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*.

[14] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 4401–4410. https://doi.org/10.1109/CVPR.2019.00453

[15] 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).

[16] 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

[17] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2019. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. *CoRR* abs/1907.11922 (2019). arXiv:1907.11922 http://arxiv.org/abs/1907.11922

[18] 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

[19] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. 2018. Are GANs Created Equal? A Large-Scale Study. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 698–707. http://papers.nips.cc/paper/7350-are-gans-created-equal-a-large-scale-study

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 8024–8035. http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library

[21] Tiziano Portenier, Qiyang Hu, Attila Szabó, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. 2018. Faceshop: deep sketch-based face image editing. *ACM Trans. Graph.* 37, 4 (2018), 99:1–99:13. https://doi.org/10.1145/3197517.3201393

[22] 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

[23] 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.

[24] Ivan E. Sutherland. 1964. Sketch pad a man-machine graphical communication system. In *DAC*.

[25] 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,*.

[26] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. 2016. Instance Normalization: The Missing Ingredient for Fast Stylization. *CoRR* abs/1607.08022 (2016). arXiv:1607.08022 http://arxiv.org/abs/1607.08022

[27] 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

[28] Martin Weber. 2018. AutoTrace. http://autotrace.sourceforge.net/

[29] 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

[30] 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

[31] Robert C. Zeleznik, Kenneth P. Herndon, and John F. Hughes. 1996. SKETCH: An Interface for Sketching 3D Scenes. In *Proc. SIGGRAPH*. 163–170.

[32] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei 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

[33] 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.