LinesToFacePhoto: Face Photo Generation From Lines With Conditional Self-Attention Generative Adversarial Network

Yuhang Li
lyh9001@mail.ustc.edu.cn
NEL-BITA, University of Science and Technology of China

Xuejin Chen*
xjchen99@ustc.edu.cn
NEL-BITA, University of Science and Technology of China

Feng Wu
fengwu@ustc.edu.cn
NEL-BITA, University of Science and Technology of China

Zheng-Jun Zha
zhazj@ustc.edu.cn
NEL-BITA, University of Science and Technology of China

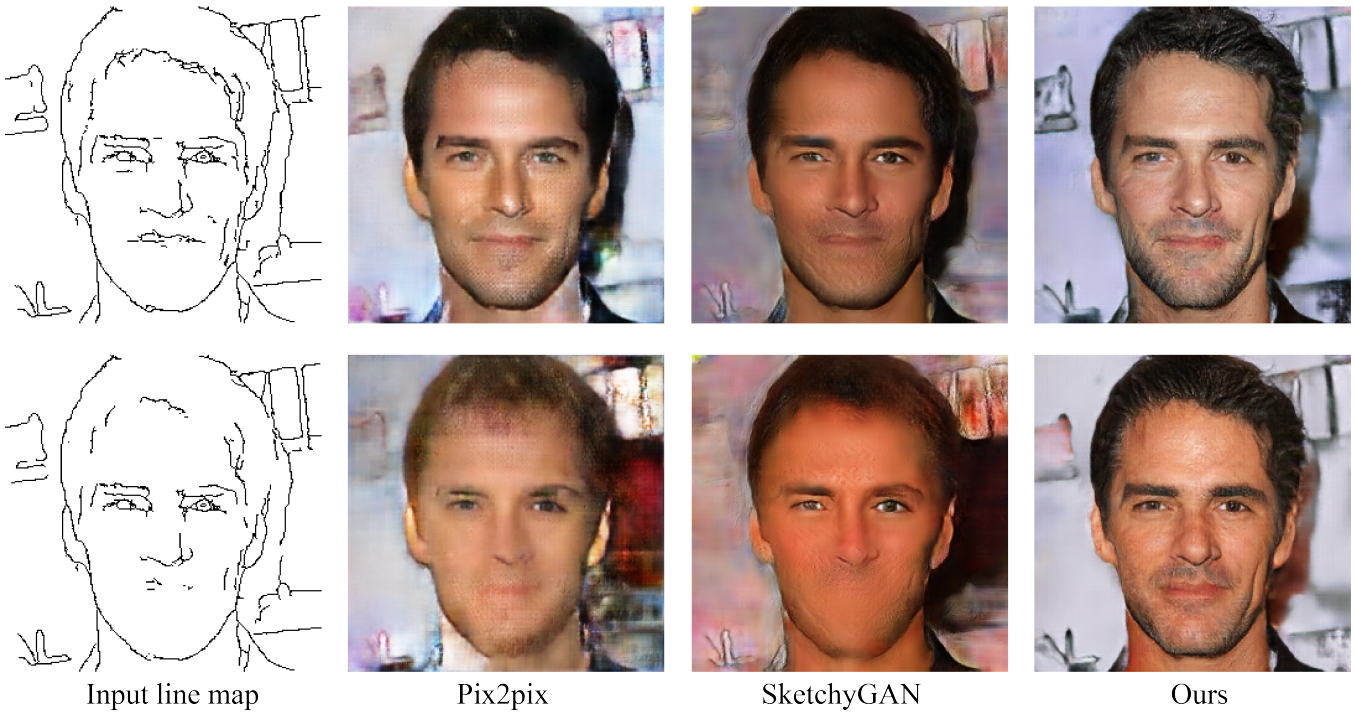

Figure 1: From sparse lines that coarsely describe a face, photorealistic images can be generated using our conditional self-attention generative adversarial network (CSAGAN). With different levels of details in the conditional line maps, CSAGAN generates realistic face images that preserve the entire facial structure. Previous works [4, 13] fail to synthesize certain structural parts (i.e. the mouth in this case) when the conditional line maps lack corresponding shape details.

ABSTRACT

In this paper, we explore the task of generating photo-realistic face images from lines. Previous methods based on conditional generative adversarial networks (cGANs) have shown their power to generate visually plausible images when a conditional image and an output image share well-aligned structures. However, these models fail to synthesize face images with a whole set of well-defined structures, e.g. eyes, noses, mouths, etc., especially when the conditional line map lacks one or several parts. To address this problem, we propose a conditional self-attention generative adversarial network (CSAGAN). We introduce a conditional self-attention mechanism to cGANs to capture long-range dependencies between different regions in faces. We also build a multi-scale discriminator. The large-scale discriminator enforces the completeness of global structures and the small-scale discriminator encourages fine details, thereby enhancing the realism of generated face images. We

*Xuejin Chen is the corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350854

evaluate the proposed model on the CelebA-HD dataset by two perceptual user studies and three quantitative metrics. The experiment results demonstrate that our method generates high-quality facial images while preserving facial structures. Our results outperform state-of-the-art methods both quantitatively and qualitatively.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**;

KEYWORDS

self-attention; conditional generative adversarial nets; face; line map; realistic images

ACM Reference Format:
Yuhang Li, Xuejin Chen, Feng Wu, and Zheng-Jun Zha. 2019. LinesToFacePhoto: Face Photo Generation From Lines With Conditional Self-Attention Generative Adversarial Network. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350854

1 INTRODUCTION

When creating something from scratch, a natural and intuitive way is to draw lines. Line drawing is an effective form of visual thought. It describes the structure and shape of a desired object more specifically than text. Turning lines into photorealistic images has drawn a lot of attention in computer graphics and computer vision for many years. Benefiting from the massive amounts of images on the Internet, many approaches in the past decade have been proposed based on line-based image retrieval and image synthesis techniques [3, 9, 29, 36]. While these methods successfully maintain the primary structure of the desired object or scene, they typically fail to generate fine details due to the limited capability of traditional image synthesis techniques.

With the emergence of deep neural networks (DNN), a series of approaches based on generative adversarial networks (GANs) have been proposed for realistic image synthesis [18, 30]. A generator and a discriminator in GANs are trained by playing a min-max game to guide the generated samples to become indistinguishable from real ones. Image-to-image translation, which is a specific application of conditional GANs, aims to translate an image in one domain to a target image in another domain, while preserving main contents and structures in these two images. Since the first image-to-image translation model, pix2pix [13], was proposed, there have been many variants in both supervised and unsupervised manners [16, 20, 33, 37, 39, 40]. These models successfully synthesize realistic textures when complete and detailed structures are given in the conditional image.

However, when the structure is partially provided in the conditional image, which is exactly the case of line drawings or edge maps, previous models fail to complete the missing structure. This is mainly because these methods strictly follow the provided edges when synthesizing the generated image; thus, they do not generate new structures at the place where few edges are provided. Since faces are composed of well-defined structural parts, e.g. noses, mouths, eyes, etc., the synthesized face images should contain the whole set of these structural parts to appear realistic, even when the conditional line maps lack edges around the supposed locations of these parts. As Figure 1 shows, using our method, the generated images from two line maps with different levels of details are more realistic because of the complete global structures and fine textures. Previous methods [4, 13] fail to render realistic face images while edges around the mouth area are incomplete.

The underlying reasons for this failure are mainly two-fold. First, existing GANs are built primarily on convolutional layers. Since the convolutional operator has a local receptive field depending on its kernel size, a large receptive field is achieved by stacking multiple convolutional layers. However, it is non-trivial for current network optimizers to discover proper parameter values that model the long-range dependence through several convolutional layers [38]. Secondly, existing discriminators used in GANs focus on examining local patches instead of capturing the global information; therefore, they fail to enforce the generator to synthesize global structure of the generated image.

Considering the first issue, we propose a conditional self-attention mechanism to the image-to-image translation model generator to address the problem. Self-attention, which computes the response at one position as a weighted sum of the features at all positions, is able to capture long-range dependency across different parts [6, 31, 34, 38]. In order to adapt the conditional setting of image-to-image translation and encourage the GAN model to fully exploit the information from the conditional image directly, we propose a *conditional self-attention module (CSAM)*, which enables the higher layers to sense information from the conditional image and capture long-range dependency. For the second issue, we establish a multi-scale discriminator to capture information from different levels. The small-scale discriminator has a local receptive field and improves the fine textures of local patches, while the large-scale discriminator ensures the completeness of the global structure in the generated images.

In this paper, we focus on the task of portrait photograph generation from line drawings, while preserving well-defined face structures, which are critically important for the realism of face photos. Our contributions are summarized as follows:

(1) We first introduce the self-attention mechanism to line-to-image translation and propose a novel conditional self-attention generative adversarial networks. Unlike convolution-based methods, the proposed model is able to model long-range dependence and global structures in face images.
(2) We show the effectiveness of the proposed model by a series of experiments on the CelebA-HD dataset. Our method generates high-quality face images from sparse lines and preserves facial structures. The proposed CSAGAN outperforms state-of-the-art methods both quantitatively and qualitatively.

2 RELATED WORK

Our CSAGAN method for generating face photos from line maps is built upon previous work on image-to-image translation, attention mechanism, and line-based image synthesis. We discuss most related techniques in this section.

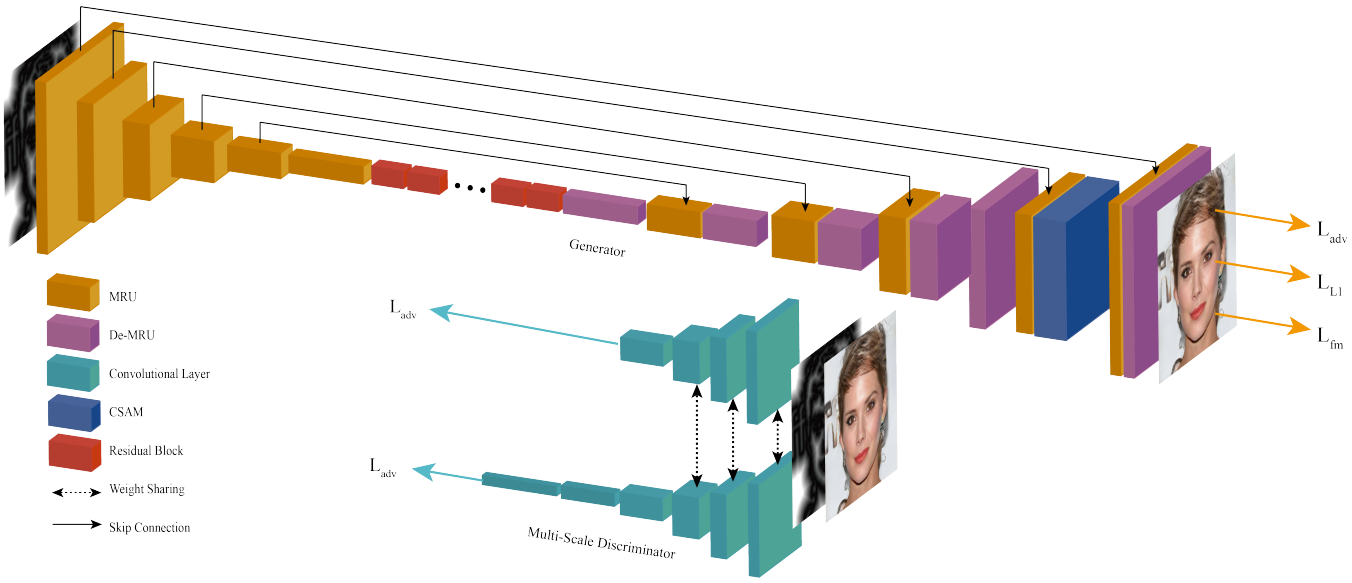

Figure 2: The architecture of our model. The proposed CSAM (the blue block) is added before the last convolutional layer. The multi-scale discriminator (only two are drawn) are applied to encourage the generator to produce realistic results with complete structure and delicate textures.

2.1 Image-to-Image Translation with GANs

Given an image in one domain, image-to-image translation methods generate a corresponding image in another domain, while depicting the same scene or object in different styles. The pix2pix method [13] first introduced image-to-image translation with conditional GAN for a range of applications. However, it is difficult for the convolution-based architecture used in pix2pix to discover long-range dependence across different regions. Moreover, the patch-wise discriminator in pix2pix can not ensure that global structures are well captured. Following pix2pix, many supervised techniques have been proposed to improve the resolution and details of target images. [15] studied how to generate images of outdoor scenes from semantic label maps coupled with attributes. [40] presented a framework that encourages the connection between the output and the latent code to be invertible so as to model the multi-modal distribution of possible outputs. [2, 33] used coarse-to-fine refinement frameworks to synthesize high-resolution photographic images from semantic label maps. In comparison, our work focuses on translating rough lines to realistic face photos, in which the key challenge is to learn the global structure and long-range dependence across different regions in a face image.

2.2 Self-Attention Mechanism

In order to sense global structures in a large receptive field, several convolutional layers and large kernel sizes have typically been required in previous GAN-based techniques. However, simply stacking convolutional layers or increasing kernel sizes seriously harms the computational efficiency. Self-attention, which computes the response at one position as a weighted sum of the features at all positions, is able to capture long-range dependence across different parts. [31] applied self-attention to capture global dependence in sequential data for machine translation, and they demonstrated the plausible effectiveness of the self-attention mechanism. [24] studied on combining the self-attention mechanism with autoregressive models, and proposed an image transformer model for image generation. Inspired by non-local operations in computer vision, [34] utilized the self-attention mechanism as a non-local operation to model long-range spatial-temporal dependence for video processing. [38] introduced self-attention to unconditional GANs and showed its advantages in generating natural images from noise vectors. [27] focused on saliency detection, utilizing a recurrent structure for shallow layers and a self-attention module for deep layers. [32] proposed a novel parallax version of self-attention for stereo image super-resolution. Inspired by previous works, we first explore the self-attention mechanism in the context of lines-to-photo translation to exploit global structures and long-range dependence between different parts in faces.

2.3 Lines-based Synthesis

Synthesizing images and models [5] from strokes or lines is not a novel idea. Earlier techniques of image synthesis from lines [3, 9] search for image patches in a large-scale database using the drawn lines, and then they fuse the retrieved image patches. With recently developed GANs, image-to-image translation techniques have been applied to the edge-to-photo task [13, 33]. However, these general frameworks, which are not specially designed for line drawings, require input edge maps that contain complete and carefully drawn lines to generate visually pleasing results. Taking hand-drawn sketches as input, SketchyGAN synthesizes plausible images for 50 object categories [4] . A masked residual unit (MRU) is proposed to improve the information flow by injecting the input image at multiple scales. However, when the conditional line maps

Figure 3: Dense distance field representation of sparse line maps.

lack specific structural parts, these GAN-based methods suffer from incomplete structures in the generated images. In comparison, our method learns long-range dependence in face images and produces photo-realistic images from line maps of different detail levels.

3 METHOD

In this section, we introduce our Conditional Self-attention Generative Adversarial Network (CSAGAN) for translating line maps to photo-realistic photos of human faces. The architecture of our model is shown in Figure 2. The generator is based on an encoder-decoder architecture with residual blocks. Skip connections [25] are applied between corresponding layers in the encoder and decoder. We adopt masked residual units (MRUs) [4] in our framework. We add our proposed conditional self-attention module (Sec.3.1) before the last MRU to model long-range dependence among feature maps. The conditional line map is resized and concatenated into feature maps at multiple scales to use as input for the MRUs and CSAM. Finally, to encourage the generator to produce realistic face images with complete structures and fine textures, we use a multi-scale discriminator (Sec. 3.2) to classify a face image globally and image patches locally as real or synthesized.

Input to CSAGAN. Since line maps are very sparse and rough, we adopt a dense representation using a distance transform. From a black-white line map, an unsigned Euclidean distance field is calculated as the conditional image. Figure 3 shows two examples of the distance fields generated from two line maps with different levels of details. Compared to the sparse and rough line maps, the dense distance fields spread the shape information to all pixels so that the extracted feature maps are more robust to incompleteness and noise in the input line maps. Similar ideas of using distance field representations can be found in several sketch-based applications [4, 11, 23]. On the other hand, some conditional GANs add a noise vector to the generator as input to avoid producing a deterministic output. However, the pix2pix model has shown that the noise vector is ignored by the generator and hardly changes the output. We observe the same phenomenon in our experiments, thus we do not apply noise vectors in our model.

3.1 Conditional Self-Attention Module (CSAM)

Inspired by SAGANs [38], we propose a conditional self-attention module (CSAM) for our lines-to-photo translation task to extract long-range dependence. This module is designed as a general module of conditional frameworks and can be added after any existing conditional modules of feature extraction. Given feature maps extracted from the previous layer $\mathbf{a} \in \mathbb{R}^{C\times H\times W}$ and a resized conditional line map $\mathbf{x} \in \mathbb{R}^{1\times H\times W}$ that matches the resolution of the current layer of feature maps. we concatenate them to get $[\mathbf{a}, \mathbf{x}]$ as conditioned features, where $[\cdot,\cdot]$ is the concatenation operation, and C, H, and W are the number of channels, height, and width of the feature map $\mathbf{a}$. This allows the network to form the attention based on the conditional image as well as the feature maps. In order to calculate the attention, we map the conditional features $[\mathbf{a}, \mathbf{x}]$ to two feature spaces:

$$f([\mathbf{a}, \mathbf{x}]) = \mathbf{W}_f[\mathbf{a}, \mathbf{x}], \tag{1}$$

$$g([\mathbf{a}, \mathbf{x}]) = \mathbf{W}_g[\mathbf{a}, \mathbf{x}], \tag{2}$$

where $\mathbf{W}_f$, $\mathbf{W}_g \in \mathbb{R}^{\hat{C}\times(C+1)}$ are trainable weights and are implemented by 1×1 convolutions. Here, we use $\hat{C} = C/8$ in our experiments following the setting of SAGAN [38].

Let $\mathbf{B} \in \mathbb{R}^{N\times N}$ be the attention map, where $N = H \times W$. Every element in $\mathbf{B}$, denoted as $b_{j,i}$, indicates the extent to which the model attends to the i^{th} pixel while synthesizing the j^{th} pixel. $b_{j,i}$ is calculated by

$$b_{j,i} = \frac{exp(s_{ij})}{\sum_{i=1}^{N} exp(s_{ij})}, \tag{3}$$

in which $s_{ij} = f([\mathbf{a}, \mathbf{x}])^T g([\mathbf{a}, \mathbf{x}])$. Next, we use $b_{j,i}$ as the attention weights and compute the response map $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_N) \in \mathbb{R}^{C\times N}$ at every position as a weighted sum of the features at all positions, where

$$\mathbf{r}_j = \sum_{i=1}^{N} b_{j,i} h([\mathbf{a}, \mathbf{x}]), \tag{4}$$

where $h([\mathbf{a}, \mathbf{x}]) = \mathbf{W}_h[\mathbf{a}, \mathbf{x}]$ and $\mathbf{W}_h \in \mathbb{R}^{C\times(C+1)}$. As suggested in [38], we further multiply the response of the attention layer by a scale parameter γ and add it back to the input feature maps. The final output is calculated by

$$\mathbf{o}_i = \gamma \mathbf{r}_i + \mathbf{a}_i, \tag{5}$$

where γ is a trainable value and is set to 0 at the beginning of the training process. In this way, the network learns local dependence at early stages in the training process, and then it learns long-range dependence by assigning more weights to the non-local evidences progressively.

3.2 Multi-Scale Discriminator

The discriminators for the pix2pix model and SketchyGAN are patch-wised, which distinguish real/fake images patch by patch convolutionally in local receptive fields that are much smaller than the size of the input images. The average value of all responses is calculated as the ultimate output of the patch discriminator. This is based on the assumption of independence between pixels separated by more than one patch's diameter. However, since the structural constraint is global information across an entire image, the patch-wise discriminator fails to capture the global structure. We design a multi-scale discriminator consisted of N_D subnetworks with different depths and, therefore, different sizes of receptive fields in their last layers. The receptive field in the last layer of the deepest subnetwork is as large as the entire image to capture the global structure. These subnetworks share weights with each other in first few layers since the lower-level features of these discriminators should be the same.

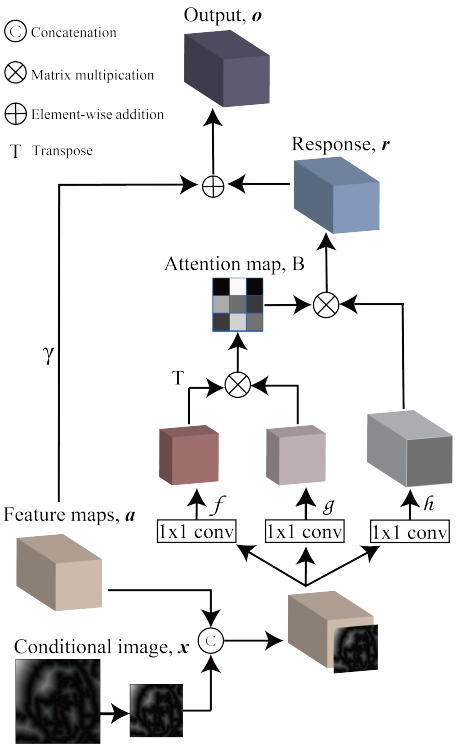

Figure 4: Components of the proposed CSAM. Given the conditional image and feature maps from the previous layer, the output feature maps are calculated in a self-attention manner.

We note that similar ideas of employing multiple discriminators has already been raised by [2, 7, 33]. They resize the real/fake images to multiple scales and apply discriminator subnetworks with the same architecture to sense different levels of structures of the real/fake images. In comparison, we fix the size of the real/fake images, and apply discriminators of different depths to achieve multiple sizes of receptive fields. It is more stable and computationally efficient to share the weights of the shallow convolutional layers. Comparison experiments with different numbers of discriminator subnetworks and previous multi-scale discriminators are described in Sec.4.5.

3.3 Loss Function

With the multi-scale discriminator D, which consists of subnetworks $\{D_i, i = 0, 1, \cdots, N_D\}$, the adversarial loss is written as

$$\begin{aligned}\mathcal{L}_{adv}(G;D) = \frac{1}{N_D}\sum_{i=0}^{N_D} E_{(\mathbf{x},\mathbf{y})\sim p_{data}(\mathbf{x},\mathbf{y})}\big[\log D_i(\mathbf{x},\mathbf{y})\big] \\ + E_{\mathbf{x}\sim p_{data}(\mathbf{x})}\Big[\log\Big(1 - D_i\big(\mathbf{x}, G(\mathbf{x})\big)\Big)\Big].\end{aligned} \tag{6}$$

Similar to the pix2pix model, we also adopt an $L1$ loss to encourage the generated image $G(\mathbf{x})$ from a line map $\mathbf{x}$ to be close to its ground truth image $\mathbf{y}$. The $L1$ loss is given by

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{data}(\mathbf{x},\mathbf{y})}\big[\|\mathbf{y} - G(\mathbf{x})\|_1\big]. \tag{7}$$

In order to achieve better perceptual quality in generated face images, we add a feature matching loss $\mathcal{L}_{fm}$. The feature matching loss [33] is a variant of perceptual loss [8, 10, 14], which aims to minimize errors between generated images and corresponding ground-truth images in the feature space. Different from the previous techniques that employ perceptual loss through a pretained VGG model, the feature matching loss uses the feature maps produced by the discriminator in our CSAGAN. This is because the VGG models used in previous methods are always trained with the ImageNet dataset and have domain gaps with face images. Our discriminator is trained specially for face images; therefore, it is more suitable for extracting features that present perceptual information of faces. Specifically, let $D_i^q(\cdot)$ be the output of the q-th layer of the i-th discriminator subnetwork with n_i^q elements, and the feature matching loss is given:

$$\mathcal{L}_{fm}(G) = \frac{1}{N_D N_Q}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{data}(\mathbf{x},\mathbf{y})}\sum_{i=0}^{N_D}\sum_{q\in Q}\frac{1}{n_i^q}\|D_i^q(G(\mathbf{x})) - D_i^q(\mathbf{y})\|_1, \tag{8}$$

where Q is the set of selected layers of discriminators, and N_Q is the number of selected layers. We select the last three convolutional layers from each discriminator subnetwork in our experiments.

By combining the multi-scale discriminator and the feature matching loss, the full objective to train our CSAGAN is

$$\min_G \max_D \mathcal{L}_{adv}(G;D) + \lambda\mathcal{L}_{L1}(G) + \mu\mathcal{L}_{fm}(G). \tag{9}$$

where λ and μ are the weights to balance the three losses. We set $\lambda = 100.0$, and $\mu = 1.0$ in our experiments.

3.4 Training Techniques

Training GANs is non-trivial because it is hard for generator and discriminator networks to find equilibria in the adversarial minmax game. We apply several techniques to stabilize our training.

Two-Timescale Update Rule (TTUR). Previous works [12, 38] suggest that separate learning rates for the generator and the discriminator are able to compensate for the problem of slow learning in the discriminator. TTUR is applied in our training process and shown to be effective.

Spectral Normalization. Spectral normalization [22] is a recently proposed normalization technique, which restricts the spectral norm in each layer of the discriminator to constrain its Lipschitz constant. Spectral normalization is computationally efficient and requires no extra hyper-parameter. Furthermore, spectral normalization is beneficial for generator training because it prevents unusual gradients [38]. We apply spectral normalization for both our generator and discriminators.

Multi-Stage Training. In order to stabilize training,, we divide the training process into three stages. In the first stage, we train the model without CSAM. Then, we add CSAM and train the CSAM while fixing the weights of the other layers in the second stage. Finally, we fine-tune the entire CSAGAN model together.

4 EXPERIMENT

We apply the proposed CSGAN framework to generate realistic photos from sparse line drawings of faces. We conducted a series of experiments to demonstrate the effectiveness of our method for preserving facial structures and generating fine details. Comparisons with other state-of-the-art methods also show the superiority of the proposed CSGAN.

Dataset. To train our network, we use the CelebA-HD dataset [16], which contains 30K high-resolution celebrity images. We randomly select 24K images for training and 6K for testing. All the images are resized to 256×256 in our experiments. To generate pairs of line drawings and face photos for supervised training, we adopt a pipeline similar to pix2pix. Specifically, edges are first extracted using a deep edge detector named Holistically-nested Edge Detector (HED) [35]. In a generated edge map, each pixel has a value p_{HED} indicating the probability of it being an edge. Several post-processing steps, including thinning, short edge removal, and erosion, are conducted to obtain simpler and clearer line maps with fewer edge fragments.

Training Parameters. We use the Adam [17] optimizer with momentum parameters $\beta_1 = 0.5, \beta_2 = 0.999$. We update one step for either G or D alternatively, and batch size is set to 8. Either the first or the second training stage lasts 100 epochs with an initial learning rate $lr_G = 0.0001$ for the generator and $lr_D = 0.0004$ for the discriminator, while the third stage lasts 50 epochs with initial learning rates $lr_G = 0.00001$ and $lr_D = 0.00004$. The learning rates decay at the halfway point of each stage. The entire training process takes about seven days on eight GeForce GTX 1080Ti GPUs.

4.1 Evaluation Metrics

The evaluation of generative models is an open and complicated task. A model with good performance with respect to one criterion does not necessarily imply good performance with respect to another criterion [21, 28]. Traditional metrics, such as pixel-wise mean-squared error, do not present the joint statistics of the synthesized samples, and therefore are not able to evaluate the performance of a conditional generated model.

Since the goal of our lines-to-image translation is to generate face images that are visually plausible, we compare the results of different models with perceptual user studies, which are commonly used for evaluating GAN models [2, 7, 19, 26, 33]. Following a similar procedure as described in [2], we conduct two kinds of experiments: an unlimited time user study and a limited time user study. In addition, we use three popular quantitative evaluation metrics, the inception score (IS) [26], Fréchet Inception Distance (FID) [12], and Kernel Inception Distance (KID) [1], which are proved to be consistent with human evaluation in assessing the realism of images. More details are explained below.

User Study with Unlimited Time. In every trial of the user study with unlimited time, we randomly select a conditional line map from the testing dataset and generate two synthesized images using two approaches. The two synthesized images are displayed randomly on the left and right side of the conditional line map. The user has unlimited time to pick the one that "is more realistic and matches the conditional image better". No feedback is provided after each trial to avoid disturbing the user's perceptual judgment and preference.

User Study with Limited Time. In the study with limited time, we evaluate how quickly users perceive the differences between images. For each line map, we obtain four corresponding face images (one ground truth image and three synthesized images generated by our method, pix2pix [13], and SketchyGAN [4], respectively). In each comparison, we randomly select two images from these four corresponding face images and show the two selected images with the conditional line map. Similarly, the two images are displayed to the user on the left and right sides of the line map randomly. Within a duration randomly selected from a set of $\{1/8, 1/4, 1/2, 1, 2, 4, 8\}$ seconds, the user is asked to pick the one that "is more realistic and matches the conditional image better." We compute the percentage of the results generated by different methods that are preferred with different time durations.

Inception Score (IS). IS [26] computes the KL divergence between the conditional class distribution and the marginal class distribution. Although it has been pointed out that IS has serious limitations because it focuses more on the recognizability of generated images rather than the realism of details or intra-class diversity [28], it is still widely used to compare the quality of generated images.

Fréchet Inception Distance (FID). FID [12] is a recently proposed and widely used evaluation metric for generative models. It is shown to be consistent with human perceptual evaluation in assessing the realism of generated samples. It employs an Inception network to extract features and calculates the Wasserstein-2 distance between features of generated images and real images. Lower FID values indicate that the synthetic distribution is closer to the real distribution.

Kernel Inception Distance (KID). Similar to the FID, KID [1] measures the difference between two sets of samples by calculating the squared maximum mean discrepancy between Inception representations. Moreover, unlike the FID, which is reported to be empirically biased, KID has an unbiased estimator with a cubic kernel [1], and it matches human perception more consistently.

4.2 Comparisons with Previous Methods

We compare the proposed model with two state-of-the-art lines-to-photo translation methods, pix2pix [13] and SketchyGAN [4]. We train the pix2pix model on our edge-face dataset using the default settings described in [13]. SketchyGAN is originally designed for multi-class sketch-to-image generation. We remove its classification branch and the loss term for classification, and train the pruned network with our line-face dataset.

Firstly, the unlimited time user study was conducted to evaluate the perceptual quality of generated images. 50 users participated in these experiments, and each user was tested with about 250 trials. The results are reported in Table 1. These results show that given unlimited time, users are able to discover the visual differences between the generated face images using our model and previous ones. Compared with pix2pix and SketchyGAN, our results are significantly preferable according to the test users.

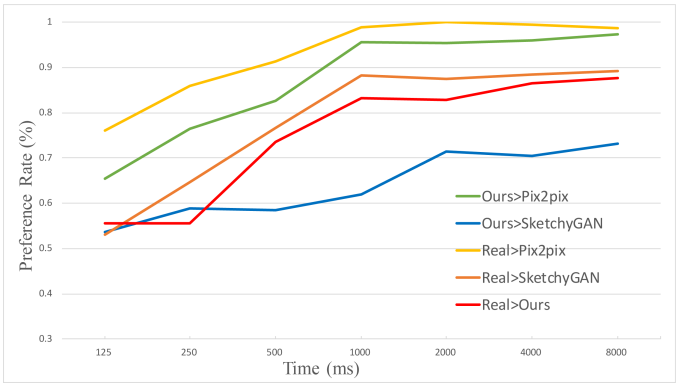

Figure 5: The results of limited time user studies. Each line indicates the user preference rate of one method over another. Users observe more differences between these methods as the display time lengthens..

Secondly, the limited time user studies were conducted to evaluate how quickly users can perceive the differences between images generated by different methods. Figure 5 shows the results. When images are shown for a very short time (1/8 seconds), the users are not able to sense the differences among different methods and the ground truth.As the time increases, more differences are perceived by users, and more users prefer the results generated by our CSAGAN.

Thirdly, Table 2 lists the quantitative comparisons of our method and others. As we can see, our full model surpasses the pix2pix model and SketchGAN by a large margin with regard to the mean values of IS/FID/KID, demonstrating our model's capability to generate more realistic face photos.

Finally, Figure 6 shows a group of synthesized face images using the proposed model and previous methods (better viewed in color). We observe that the results of our model contain more details, especially in the areas with hairs, whiskers, and highlighted regions, while the results of the previous models are over-smoothed and lack realistic details. Moreover, our results appear more realistic with regard to the illumination of faces.

Table 1: Results of user study with unlimited time.

	Pix2pix [13] vs Ours	SketchyGAN[4] vs Ours
Preference	6.9%/93.1%	24.0%/76.0%

4.3 Ablation Study

We examine the importance of every component within our model based on IS/FID/KID, shown in the third and fourth row of Table 2. The experiments were conducted by removing each specific part from the full model and then training the rest of th model without the absent part. Specifically, we remove 1) the proposed CSAM (ours w/o CSAM), and 2) the multi-scale discriminator, using only the patch discriminator D_p (ours w/ D_p). As we can see, the performance of our model without CSAM drops dramatically compared to the full model, indicating the critical importance of CSAM in our model. The performance of our model also benefits from the multi-scale discriminator.

We also visualize the attention maps to demonstrate how pixels in different locations are related and dependent in the learned

Table 2: Quantitative comparison. (Larger IS and lower FID/KID represent better results.)

Models	IS	FID	KID
Pix2pix [13]	2.55 ± 0.20	605.97 ± 13.95	3.05 ± 0.08
SketchyGAN [4]	2.75 ± 0.17	479.09 ± 15.24	2.11 ± 0.09
Ours w/o CSAM	2.65 ± 0.08	426.78 ± 17.23	1.62 ± 0.04
Ours w/ D_p	2.73 ± 0.09	413.26 ± 13.92	1.57 ± 0.07
Ours w/ $D_{pix2pixHD}$	2.70 ± 0.08	398.78 ± 14.60	1.40 ± 0.05
Ours, $N_D = 2$	2.71 ± 0.13	332.00 ± 9.26	0.98 ± 0.04
Ours, $N_D = 3$	**2.78 ± 0.09**	269.96 ± 8.18	0.63 ± 0.04
Ours, $N_D = 4$	2.76 ± 0.10	**269.59 ± 8.89**	**0.62 ± 0.05**

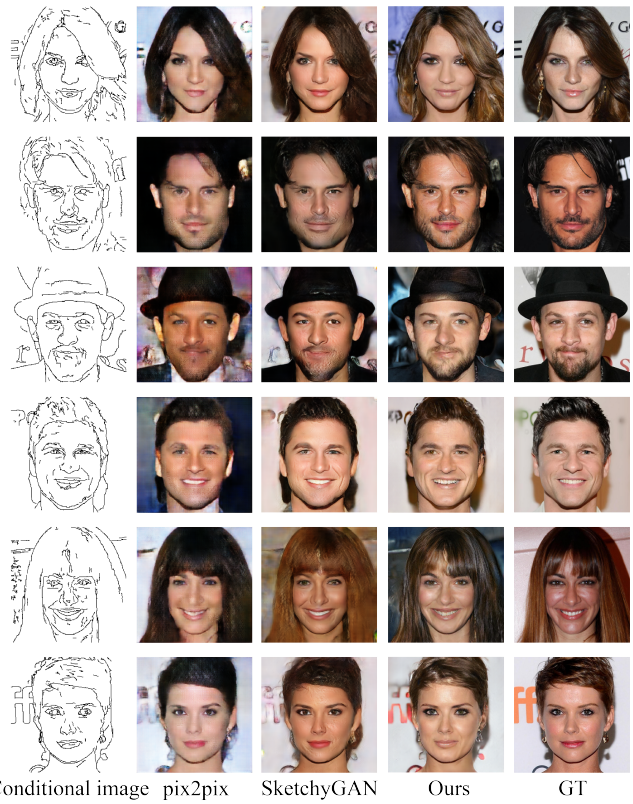

Figure 6: The above face images are generated from edge maps using three methods: pix2pix [13], SketchyGAN [4] (without a classification network), and our proposed method. Ground truth (GT) images that we use to obtain the edge maps are shown in the right-most column. The results generated by our model contain more details, especially in the areas with hairs, whiskers, and highlighted regions. Also, our results appear more realistic with regard to the illumination of faces.

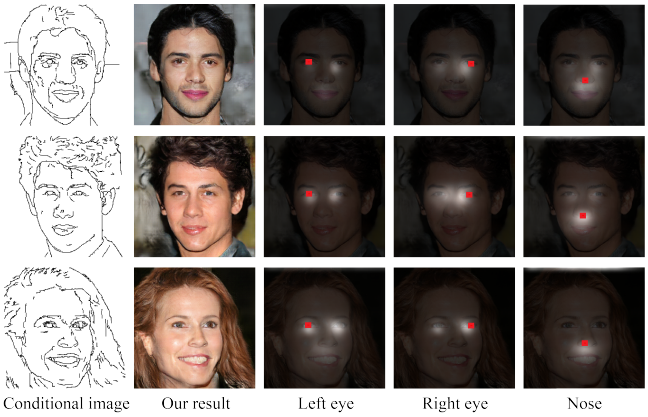

Figure 7: The attention maps are shown with the conditional edge maps and the generated images. Three locations of the nose and two eyes are marked in red while the attention maps with respect to these locations are shown. The larger values in the attention maps are brighter in the figure. We observe that long-range dependence between different parts of faces is captured by our CSAM.

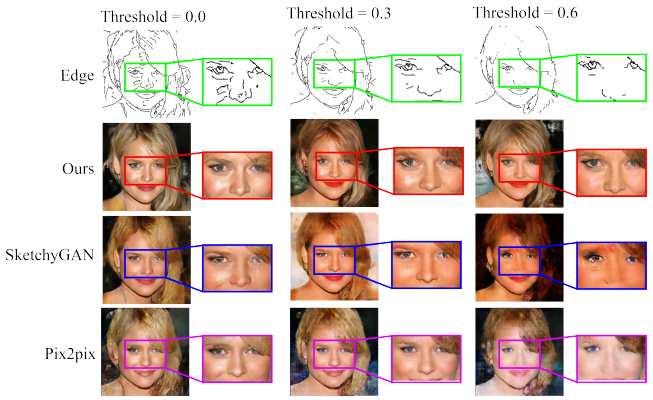

Figure 8: The face images generated from line maps with different levels of details. Our proposed model is able to generate realistic face images with complete structures and fine textures (the second row). Whereas, the two previous models [4, 13] fail to generate the nose and the left eye when τ is set to 0.3 and 0.6. The area around the nose is zoomed in, and the ground truth is displayed on the right.

model. Figure 7 shows a group of examples of attention maps. Three locations (i.e., the nose and the two eyes) are marked in red, and the attention maps with respect to these locations are shown, respectively. The larger values in the attention maps are brighter in the figure. We observe that the long-range dependence is captured by the CSAM. For example, to generate the pixels in one eye, the regions of both eyes are assigned high attentions. In another words, the information for generating a specific pixel comes from not only its local area but also related regions far away from this pixel.

4.4 Different Levels of Details in Line Maps.

To evaluate the robustness of our CSAGAN, we use line maps with different levels of details to produce face images. As discussed in the dataset construction, we produce edge maps based on p_{HED} with several post-processing steps. The lines in each edge map are generated by keeping edge pixels that are $p_{HED} > \tau$. A larger τ value causes less detail in the edge maps. By setting different values for τ, we generate edge maps with different levels of details, as shown in Figure 8. The proposed model is robust enough to generate face images with the whole structure when the inputs are line maps with different levels of details. In comparison, the two previous models fail to generate some parts of the face (i.e., the nose) when detail edges are missing in the line maps with larger τ value (0.3 and 0.6 in this case).

4.5 Comparison of Different Multi-scale Discriminators

We compare our multi-scale discriminator with its variants and the one from previous work [33]. More specifically, if we let N_D be the number of discriminator subnetworks, we train the models with the same generator and discriminator as $N_D = \{2, 3, 4\}$ subnetworks, denoted as ours, $N_D = \{2, 3, 4\}$. Each subnetwork shares weights with others in the first few layers, while the depths are different. Therefore, the receptive fields of the last layers in the subnetworks are different, and the subnetworks distinguish generated samples from real ones in different scales. Quantitative comparison results based on IS/FID/KID are shown in the last three rows of Table 2.

Also, we compare our multi-scale discriminator with the one in [33]. Specifically, we use our generator and switch our discriminator to the one from [33] ($N_D = 3$) and train this model with our three-stage training process, which is denoted as ours w/ $D_{pix2pixHD}$. Results shown in Table 2 indicate that our multi-scale discriminator (ours, $N_D = 3$) exceeds its counterpart in the measure of IS/FID/KID and shows its advantages on quantitative evaluation.

5 CONCLUSION

In this work, we propose a conditional self-attention GAN (CSAGAN) to synthesize photo-realistic face images from sparse lines. By introducing the self-attention mechanism and a multi-scale discriminator into conditional GANs, our method is able to capture long-range dependence across different regions and global structures in face images. Comprehensive experiments illustrate the effectiveness of the proposed method via two perceptual studies and three quantitative metrics. Our framework shows its promising capability to generate high-quality face images by synthesizing complete facial structures as well as fine details, even when some parts of the input line map are missing.

ACKNOWLEDGMENTS

This work was supported by the National Key Research & Development Plan of China under Grant 2016YFB1001402, the National Natural Science Foundation of China (NSFC) under Grants 61632006, 61622211, and 61620106009, as well as the Fundamental Re-search Funds for the Central Universities under Grants WK3490000003 and WK2100100030.

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*.

[2] Qifeng Chen and Vladlen Koltun. 2017. Photographic Image Synthesis with Cascaded Refinement Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 1520–1529.

[3] 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.

[4] 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.

[5] Xuejin Chen, Sing Bing Kang, Ying-Qing Xu, Julie Dorsey, and Heung-Yeung Shum. 2008. Sketching Reality: Realistic Interpretation of Architectural Designs. *ACM Trans. Graph.* 27, 2 (May 2008), 11:1–11:15.

[6] Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long Short-Term Memory-Networks for Machine Reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 551–561.

[7] Emily L Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. 2015. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In *Advances in Neural Information Processing Systems 28*. 1486–1494.

[8] Alexey Dosovitskiy and Thomas Brox. 2016. Generating Images with Perceptual Similarity Metrics based on Deep Networks. In *Advances in Neural Information Processing Systems 29*. 658–666.

[9] 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.

[10] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2414–2423.

[11] Xiaoguang Han, Zhen Li, Haibin Huang, Evangelos Kalogerakis, and Yizhou Yu. 2017. High-Resolution Shape Completion Using Deep Neural Networks for Global Structure and Local Geometry Inference. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 85–93.

[12] 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.

[13] 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.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *European Conference on Computer Vision*. 694–711.

[15] 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

[16] 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*.

[17] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015,*.

[18] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014,*.

[19] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 105–114.

[20] Ming-Yu Liu and Oncel Tuzel. 2016. Coupled Generative Adversarial Networks. In *Advances in Neural Information Processing Systems 29*. 469–477.

[21] 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.

[22] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral Normalization for Generative Adversarial Networks. In *6th International Conference on Learning Representations, ICLR 2018,*.

[23] Duc Thanh Nguyen, Binh-Son Hua, Minh-Khoi Tran, Quang-Hieu Pham, and Sai-Kit Yeung. 2016. A Field Model for Repairing 3D Shapes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5676–5684.

[24] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image Transformer. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*. 4052–4061.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*, Vol. 9351. 234–241.

[26] 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.

[27] Fengdong Sun, Wenhui Li, and Yuanyuan Guan. 2018. Self-attention recurrent network for saliency detection. *Multimedia Tools and Applications* (17 Sep 2018).

[28] 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,*.

[29] Daniyar Turmukhambetov, Neill D.F. Campbell, Dan B Goldman, and Jan Kautz. 2015. Interactive Sketch-Driven Image Synthesis. *Comput. Graph. Forum* 34, 8 (Dec. 2015), 130–142.

[30] Aäron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals, and Alex Graves. 2016. Conditional Image Generation with PixelCNN Decoders. In *Advances in Neural Information Processing Systems 29*. 4790–4798.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*. 6000–6010.

[32] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. 2019. Learning Parallax Attention for Stereo Image Super-Resolution. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 12250–12259.

[33] 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.

[34] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-Local Neural Networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 7794–7803.

[35] Saining Xie and Zhuowen Tu. 2015. Holistically-Nested Edge Detection. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 1395–1403.

[36] Kun Xu, Kang Chen, Hongbo Fu, Wei-Lun Sun, and Shi-Min Hu. 2013. Sketch2Scene: Sketch-based Co-retrieval and Co-placement of 3D Models. *ACM Trans. Graph.* 32, 4 (July 2013), 123:1–123:15.

[37] 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.

[38] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-Attention Generative Adversarial Networks. In *Proceedings of the 36th International Conference on Machine Learning*. 7354–7363.

[39] 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.

[40] 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.