Edge2photo

YUHANG LI, University of Science and Technology of China, China
XUEJIN CHEN, University of Science and Technology of China, China
SIYU HU, University of Science and Technology of China, China
ZHENG-JUN ZHA, University of Science and Technology of China, China
SING BING KANG, Microsoft Research, USA

We explore the task of translating face images from the corresponding edge maps, a specific version of image-to-image translation which has drawn a lot of interest recently. The previous proposed image-to-image translation model (pix2pix) has shown to be powerful in generating visually plausible images from the conditional images. However, in some cases this model is not able to synthesize faces with the whole set of well-defined structures, e.g. eyes, noses, mouths, and etc., especially when the conditional edge map lack of one or several parts of the structure. To address this problem, we propose conditional self-attention generative adversarial networks (CSAGANs) to capture the long-range dependencies and global structure information across images. This model is also able to leverage the information of conditional image directly. We demonstrate the effectiveness of the proposed model with experiments on translating faces of CelebA dataset from the corresponding edge maps. We evaluate our model by two kinds of perceptual user studies and Fréchet Inception Distance (FID), and show that this model xxxx (discuss words to describe the results properly).

CCS Concepts: • **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

Additional Key Words and Phrases: Generative adversarial nets, edge maps, realistic images

ACM Reference Format:
Yuhang Li, Xuejin Chen, Siyu Hu, Zheng-Jun Zha, and Sing Bing Kang. 2010. Edge2photo. *ACM Trans. Web* 9, 4, Article 39 (March 2010), 15 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Realistic image synthesis has been a hot topic in computer vision and computer graphics for years. Traditional methods [9, 10, 15] establish databases of existing images, and generate images by matching and merging images in the database patch-wisely. With the emergence of deep neural networks (DNN), several promising DNN-bsed approaches for image synthesis have been proposed. Variational autoencoders (VAEs) [21], which maximize a variational lower bound on the log-likelihood of the training data, have brought some progress in generating visually plausible images, but the generated samples suffer from being blurry. Autoregressive models [39] generate

Authors' addresses: Yuhang Li, University of Science and Technology of China, xx Rd, Hefei, Anhui, 230027, China, lyh9001@mail.ustc.edu.cn; Xuejin Chen, University of Science and Technology of China, xx Rd, Hefei, Anhui, 230027, China, xjchen99@ustc.edu.cn; Siyu Hu, University of Science and Technology of China, xx Rd, Hefei, Anhui, 230027, China, sy891228@mail.ustc.edu.cn; Zheng-Jun Zha, University of Science and Technology of China, xx Rd, Hefei, Anhui, 230027, China, xxx@ustc.edu.cn; Sing Bing Kang, Microsoft Research, xx Rd, xxx, xxx, xxxx, USA, SingBing.Kang@microsoft.com.

1559-1131/2010/3-ART39 $15.00
https://doi.org/0000001.0000001

image pixel by pixel and are able to generate convincing samples. However, it is computational inefficient to sample images from these models.

Generative adversarial networks (GANs) [12] offer a new and promising mechanism to generate images, which take noise vectors as input and train two networks playing minmax game to guide the generated samples to be indistinguishable from the real ones. Conditional GANs, which generate image from assigned conditional information instead of noise vector, are conditional versions of GANs. Conditional GANs are trained in a supervised manner and shown to be powerful in modeling the conditional distributions with respect to the assigned conditions. A variety of conditions have been applied to conditional GANs, such as discrete class labels [29], texts [47, 48], and images. Among these conditional image generation methods, image-to-image translation has drawn a lot of attention recently, which aims to apply a conditional image in one domain to generate the corresponding target image in another, reserving shared concepts, objects or scenes in these two images. Since the first image-to-image model (pix2pix) [17] was proposed, there have been many variants of this approach in both supervised and unsupervised manner [26, 45, 50, 51]. However, the pix2pix model has troubles in some cases of translating face images from the corresponding edge maps. For example, this model occasionally fails to synthesize faces with the whole set of well-defined structures, e.g. eyes, noses, mouths, and etc., especially when the conditional edge map lack of one or several parts of the structure.

The reasons behind this might be two-folded. 1) The pix2pix model is built based on convolution layers. Since the convolution operator has a local receptive field depending on the size of its kernels, a large receptive field is achieved by cooperation of several convolution layers. It is hard for the optimizer to discover parameter values that model the long-range dependencies through several convolutional layers [46]. 2) The discriminator used in the pix2pix model [17] focuses on examining local patches instead of capturing the global information, and therefore fails to guide the generator to synthesize the global structure of the conditional image.

Considering the first reason, we introduce a conditional self-attention mechanism to the generator of image-to-image models to address the problem. Self-attention [6, 40, 42, 46], which computes the response at a position as a weighted sum of the features at all positions, is able to capture the long-range dependencies across different regions of images and feature maps. In order to adapting the conditional setting of image-to-image translation and encouraging the model to leverage the information of the conditional image directly, we propose a conditional self-attention module (CSAM) which enables the higher layers to sense the conditional image. For the second reason, we consider to establish multiple discriminators to capture information of different levels, both patch-wisely and globally. We note that similar idea of multiple discriminators has been raised by [4, 7, 16, 48] who resizes the real/fake samples and applies multiple discriminators to these multi-scale samples.

In this research, we propose Conditional Self-attention Generative Adversarial Networks (SC-GANs), which translate images from one domain to another being able to capture long-range dependencies and reserve the global structures. With the help of the novel CSAMs, the conditional image is able to guide the higher layers in the architecture directly.

Our contributions are summarized as follow:

i) We firstly introduce the self-attention mechanism to image-to-image translation and propose a novel conditional self-attention generative adversarial networks for the image-to-image translation task. Unlike convolutional-based methods, the proposed model is able to model the long-range dependencies and global structure across images.

ii) We propose a multi-level discriminator to the image-to-image translation. The proposed discriminator is able to capture the global structure information as well as the local realism.

iii) We show the effectiveness of the proposed model by experiments. Two kinds of user studies are investigated to show the perceptual evaluation of the results generated by the proposed method. Quantitative evaluation is conducted by calculating the FID of the pix2pix model and the proposed model.

The rest of this article is organized as follow. Related works are presented in Section **??**. The method we proposed is introduced in Section 3. We demonstrate the effectiveness of method by experiments in Section 4. Section **??** summarizes our conclusions and outlines possible future work.

2 RELATED WORK

Our work is based on image-to-image translation frameworks, which are variants of GANs in a conditional setting. In this section, we present related research in GANs, conditional GANs, and image-to-image translation models. We also give a brief review on recently proposed attention models.

2.1 Generative Adversarial Networks (GANs)

Generative adversarial networks (GANs) [12] have obtained a great success in recent years. Based on the minmax game theory, a classical architecture of GANs contains a generator network and a discriminator network. The task of the generator take a noise vector as input and generate samples indistinguishable from the real ones, while the discriminator, in opposite, attempt to find out whether its input is real or synthesized. The minmax game played by these two networks guides the generated distribution to be similar to the real data distribution. Compared to other deep framework of image generation [21, 39], GANs are able to synthesize images with less blurriness and provide a more efficient process to generate samples. However, GANs suffer from several problems in the early stage, such as the instability of training and the mode collapse problem. To stabilize the training of GANs and enable GANs to generate images with high quality and large diversity, many efforts have been made. Deep convolutional GANs (DCGANs) [35] first introduced a convolutional architecture which led to improved visual quality. [37] proposed an approach to train discriminator in a semi-supervised fashion, granting the discriminator's internal representations knowledge of the class structure of (some fraction of) the training data it is presented. Energy based GANs (EBGANs) [49] were proposed as a class of GANs that aims to model the discriminator as an energy function. This variant converges more stably and is both easy to train and robust to hyper-parameter variations. Wasserstein distance, which acts as a loss as well as a measure of convergence in training process, is brought to GANs by [1, 13] to benefit both the stability and mode coverage. Several other works [3, 22, 28] also make progress in stabilizing the training and increasing the diversity of the results of GANs. A recently proposed work introduced self-attention mechanism to the unconditional GANs and achieved state-of-the-art results. Our work is based on GANs in a conditional setting.

2.2 Conditional Generative Adversarial Networks

Conditional GANs are generalized versions of GANs in a conditional setting. Instead of taking a noise vector as input, conditional GANs generating images based on the assigned conditions, modeling the conditional distribution of the samples. Conditional GANs were firstly introduced by [29] who treated the conditional generation problem as the inverse processing of image classification and used discrete labels as condition to generate images. Previous works have explored GANs generating images based on a wide variety of conditions. [5, 33] took both noise vectors and discrete class label as input and added a classifier task to the discriminator in two different architectures to generate images high recognizability. [8] trained convolutional networks to generate images of objects given object style, viewpoint and color. With the experiments of interpolating viewpoints, they showed that networks learn a meaningful representation of 3D models. [47, 48] generated high-resolution photo-realistic images conditioned on text descriptions in two stages, where GANs sketch the basic shapes and colors in the first stage and add details in to the generated images in the second stage. A recently proposed method[31] leveraged the conditional information in a novel way, where the discriminator involves an inner product term between the condition vector and the feature vector in a middle level layer. This formulation is based on the observation that the loss function of GANs are able to be decomposed into the sum of two log likelihood ratios. Our work utilize GANs in a conditional setting to generate images from images, which utilize the condition information directly even in the high-level layers.

2.3 Image-to-image translation with GANs

Given an image in one domain, image-to-image translation methods generate a corresponding image in another. These two images are possible representations of the same scene or object. Image-to-image translation with GANs is a special case of conditional GANs where images are applied to be conditions.

The pix2pix method [17] firstly introduced the concept of image-to-image translation. Pix2pix is train in a supervised manner, where the training dataset is a set of paired images. Pix2pix applies skip connections [36] between mirrored layers in the generator to make sure low-level information pass through its encoder-decoder architecture and uses patch discriminators to increase the performance of the generator. However, the convolution-based architecture makes it difficult to discover the long-range dependencies across the images and feature maps, and the patch-wise discriminator is not able to ensure the global structure information to be well capture by the model. In addition to pix2pix, many image-to-image tasks are trained in a supervised manner. [41**?**] used coarse-to-fine refinement frameworks to synthesis photographic images from semantic label maps. [18] studied the generating images of outdoor scenes from semantic label maps coupled with attributes. [51] presented a framework that is able to model the multi-modal distribution of possible outputs. Image-to-image translation has also been well-studied in an unsupervised setting. [25] studied on unpaired image-to-image translation by training a two-branch GAN. Each branch is composed with a encoder, a generator and a discriminator. With the idea that high-level representation of a pair of corresponding images in two domains should be the same, high-level layers share weights between two branches in encoders, generators and discriminators. CycleGAN [50], DiscoGAN [19] and DualGAN [45] developed similar architectures to translate unpaired images which contain, for each, two generators and two discriminators. These methods learn two mappings in an adversarial training process such that an input image in one domain is mapped to a generated image in another, and then the generated image is mapped to a reconstructed image which is closed to the input image in some measures. These methods shared the same idea that since the generated image is able to reconstruct the input image, it should contain the content of the input image.

Our work focuses on translating face images from corresponding edge maps in a supervised setting, which is able to learn the long-range dependencies and global structure across image.

2.4 Attention Mechanism

The convolution operation has a local receptive field. Several layers and large kernel sizes are required to sense the global structure in a large receptive field, which, however, loses the computational and statistical efficiency. Recently, attention mechanisms have been introduced to capture global dependencies [2, 44]. Self-attention [42, 46] has been shown to be powerful in a variety of tasks. [40] applied self-attention to machine translation models, and demonstrated the plausible effectiveness of self-attention mechanism. [34] studied on combining the self-attention mechanism and autoregressive models, and proposed an image transformer model in image generation. Inspired

(a) edge maps (b) GT (c) pix2pix (d) ours

Fig. 1. example

by non-local operation in computer vision, [42] utilize self-attention mechanism as a non-local operation to model long-range spatial-temporal dependencies for video processing. [46] introduced self-attention to unconditional GANs and achieved state-of-the-art results in generating natural images from noise vectors. Inspired by previous works, we explore the self-attention mechanisms in the context of image-to-image translation.

3 METHOD

In this research, we propose Conditional Self-attention Generative Adversarial Networks (CSAGANs), which translate images from one domain to another being able to capture long-range dependencies and reserve the global structures across image. We first review the pix2pix model as our baseline (Sec.3.1). And then we introduce the Conditional Self-attention Module (SCAM) (Sec. 3.2). Finally, we describe the idea of multiple level patch discriminator (Sec.3.3 and the architecture we proposed (Sec. 3.4).

3.1 The Pix2pix Model

Since our model is based on the pix2pix model [17], we review this model in this sub-section. The pix2pix model is an image-to-image translation framework based on conditional GANs, which trains a generator network G and a discriminator network D alternatively. The generator G takes a conditional image as input and outputs the corresponding target image, while the discriminator D distinguishes real images from the synthesized ones. To train these two networks in a supervise manner, a set of corresponding image pairs $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ is required as training set, where $\boldsymbol{x}_i$ is a conditional image and $\boldsymbol{y}_i$ is the corresponding target image. These two networks play a minmax game to guide the generator to model the conditional distribution of real images given the conditional images. The objective is given by:

$$\min_G \max_D \mathcal{L}_{adv}(G, D) + \lambda \mathcal{L}_{L1}(G), \tag{1}$$

where G aims to minimize this objective while D tries to maximize it as an adversarial. The adversarial loss function is generally given by

$$\mathcal{L}_{adv}(G, D) = E_{(\boldsymbol{x},\boldsymbol{y})\sim p_{data}(\boldsymbol{x},\boldsymbol{y})}[\log D(\boldsymbol{x}, \boldsymbol{y})] + E_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log(1 - D(\boldsymbol{x}, G(\boldsymbol{x})))], \tag{2}$$

and the L_1 loss is given by

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p_{data}(\boldsymbol{x},\boldsymbol{y})}[\|\boldsymbol{y} - G(\boldsymbol{x})\|_1] \tag{3}$$

The generator of the pix2pix model is a fully-convolution-based U-Net [36]. The the input of the generator is only applied to the first layer. The discriminator of the pix2pix model is a patch-wise discriminator, which examines only a patch of its input image and uses the average of outputs from all patches of the input image as the ultimate output. The size of each patch is set to 70×70. The conditional image is concatenated channel-wisely to the synthesized image or real image as the input of the discriminator.

However, in the task of translating a face image from the corresponding edge map, the pix2pix model has troubles to generate realistic face images in some cases with structural constrains. Since faces have well-defined structural parts, e.g. noses, mouths, eyes and etc., the synthesized face images should contain the whole set of these structural part to be realistic, even when the conditional edge maps lack of edges on the supposed locations of these parts. The pix2pix fails to generate realistic structural part in this circumstance. An example is displayed in Figure 1. In this example, an edge map of a face, shown on the left of the figure, only captures a part of edges of the left eye (in the green square) rather than the whole set of edges of the entire left eye. The face image generated by the pix2pix model on the condition of on this edge map is shown on the right in the figure. According to the global structural information, there should be a left eye in the red square obviously. However, we can observe that the pix2pix model fails to render a recognizable left eye in the synthesized face image.

This phenomenon might be caused by two reasons. 1) The pix2pix model is a convolution-based model which relies on convolutional operations to model the dependencies across different regions of images and feature maps. Convolutional operations have local receptive fields depending on the kernel sizes and are not able to balance between ability to model long-range dependencies and efficiency of computation and statistics in some cases [46]. 2) The discriminator used by the pix2pix model is patch-wised, based on the assumption that pixels separated by more than a patch diameter are independent to each other. This assumption is true in some cases like texture generation and style transfer, and has been applied in previous work [11, 24]. However, this assumption fails in the case with global structural constrains. Therefore the patch-wise discriminator fails to grasp the global structure information and is not able to guide the generator to be aware of the structure of the faces.

To address the problem caused by the first reason, we introduce self-attention to the generator of image-to-image models to address the problem. Self-attention [6, 40, 42, 46], which computes the response at a position as a weighted sum of the features at all positions, is able to capture the long-range dependencies across different regions of images and feature maps. In order to adapting the conditional setting of image-to-image translation and encouraging the model to leverage the information of the conditional image directly, we propose a conditional self-attention module (CSAM), which enables the higher layers to sense the conditional image, as a general module of networks. For the second reason, we consider to establish a multi-level discriminator to capture the information of its input image both patch-wisely and globally. We note that similar ideas of multiple discriminators have been raised by [4, 7, 16, 48] with different architectures. We describe CSAM and the multi-level discriminator in next sections.

3.2 Conditional Self-Attention Module (CSAM)

We improve the pix2pix model by utilizing self-attention mechanism to capture the long-range dependencies of images and feature maps. A recently proposed method [46] has introduced self-attention to unconditional GANs and achieved state-of-the-art results. Inspired by this method,

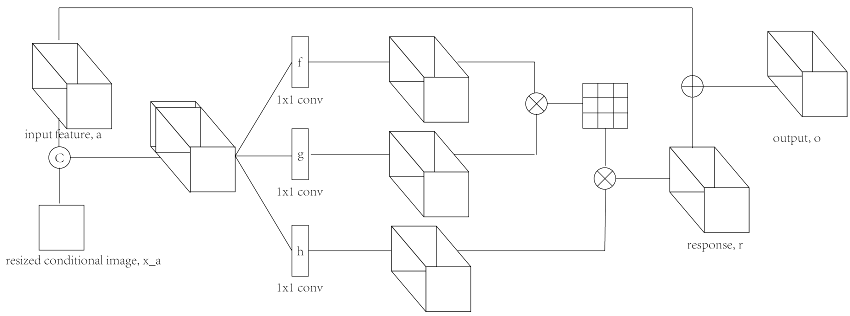

Fig. 2. CSAM

we propose a conditional self-attention module (CSAM) which is suitable for image-to-image translation framework and able to leverage the conditional information directly. This module is designed as a general module of conditional frameworks and can be added after any existing modules. We will provide details of our architecture in Subsection 3.4. The formulation of CSAM is described blow.

Given the conditional image $\boldsymbol{x} \in \mathbb{R}^{3\times N_{\boldsymbol{x}}}$ and feature maps from the previous layer $\boldsymbol{a} \in \mathbb{R}^{C\times N_{\boldsymbol{a}}}$, we first resize the conditional image $\boldsymbol{x}$ to match the size of $\boldsymbol{a}$ and get $\boldsymbol{x_a} \in \mathbb{R}^{3\times N_{\boldsymbol{a}}}$. Here $N_{\boldsymbol{x}} = H_{\boldsymbol{x}} \times W_{\boldsymbol{x}}$, where $H_{\boldsymbol{x}}, W_{\boldsymbol{x}}$ are the height and width of the conditional image $\boldsymbol{x}$. $N_{\boldsymbol{a}}$ is defined similarly for $\boldsymbol{a}$. Then we concatenate the resized conditional image $\boldsymbol{x_a}$ to the feature maps $\boldsymbol{a}$ to get $[\boldsymbol{a}, \boldsymbol{x_a}]$ as conditioned features, where $[\cdot,\cdot]$ is the concatenation operation. This allows the information of conditional image to convey to every attention module and guide the network to form the attention directly based on the conditional image.

In order to calculate the attention, we map the conditional features $[\boldsymbol{a}, \boldsymbol{x_a}]$ to two feature spaces by:

$$f([\boldsymbol{a}, \boldsymbol{x_a}]) = \mathbf{W}_f[\boldsymbol{a}, \boldsymbol{x_a}], \tag{4}$$

$$g([\boldsymbol{a}, \boldsymbol{x_a}]) = \mathbf{W}_g[\boldsymbol{a}, \boldsymbol{x_a}], \tag{5}$$

where $W_f, W_g \in \mathbb{R}^{\hat{C}\times(C+3)}$ are trainable weights and are implemented by 1×1 convolutions. Here, we use $\hat{C} = C/8$ in our experiments following the setting of previous work [46]. Let $\mathbf{B} \in \mathcal{R}^{N_a\times N_a}$ be the attention map. Every element in $\mathbf{B}$ is denoted as $b_{j,i}$ which indicates the extent to which the model attends to the i^{th} location when synthesizing the j^{th} region and is calculated by

$$b_{j,i} = \frac{exp(s_{ij})}{\sum_{i=1}^{N_a} exp(s_{ij})} \tag{6}$$

where $s_{ij} = f([\boldsymbol{a}, \boldsymbol{x_a}])^T g([\boldsymbol{a}, \boldsymbol{x_a}])$. Next, we use $b_{j,i}$ as the attention weights and compute the response $\boldsymbol{r} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \dot{;} \boldsymbol{r}_{N_{\boldsymbol{a}}}) \in \mathbb{R}^{\times N_{\boldsymbol{a}}}$ at every position as a weighted sum of the features at all positions, where

$$\boldsymbol{r}_j = \sum_{i=1}^{N_a} b_{j,i} h([\boldsymbol{a}, \boldsymbol{x_a}]), \tag{7}$$

where $h([\boldsymbol{a}, \boldsymbol{x_a}]) = \mathbf{W}_h[\boldsymbol{a}, \boldsymbol{x_a}]$ and $\mathbf{W}_h \in \mathbb{R}^{(C+3)\times(C+3)}$. As suggested in [46], we further multiply the response of the attention layer by a scale parameter γ and add back to the input feature maps. The final output is calculated by

$$\boldsymbol{o}_i = \gamma \boldsymbol{r}_i + \boldsymbol{a}_i, \tag{8}$$

where γ is trainable value and is set to 0 at the beginning of the training process. This is because at the early stage of training process, the networks are able to learn the local dependencies, and then learn the long-range dependencies by assign more weight to the non-local evidence progressively.

3.3 Multi-Level Patch Discriminator

The discriminator of the pix2pix model is patch-wised, which distinguishes the real/synthesized images patch by patch convolutional with in a local receptive field much smaller than the size of the input images. The average value of all responses is provided as the ultimate output of d. This is based on the assumption of independence between pixels separated by more than a patch diameter. However, since the structural constrain is global information across the entire image, the patch-wise discriminator may have troubles to capture this global information. We add another global discriminator D_g with a receptive field as large as the entire image to capture the global structure information. The patch discriminator D_p and the global discriminator D_g share weights in first few layers since the lower features of these discriminators should be the same, as shown in Figure 3. The objective of the minmax game therefore is modified from Equation 1 to

$$\min_{G} \max_{D_g, D_p} \mathcal{L}_{adv}(G; D_g, D_p) + \lambda \mathcal{L}_{L1}(G), \tag{9}$$

where the adversarial loss is given by

$$\begin{aligned} \mathcal{L}_{adv}(G; D_g, D_p) = & E_{(\boldsymbol{x},\boldsymbol{y})\sim p_{data}(\boldsymbol{x},\boldsymbol{y})}[\log D_g(\boldsymbol{x},\boldsymbol{y}) + \log D_p(\boldsymbol{x},\boldsymbol{y})] \\ & + E_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log(1 - D_g(\boldsymbol{x}, G(\boldsymbol{x}))) + \log(1 - D_p(\boldsymbol{x}, G(\boldsymbol{x})))], \end{aligned} \tag{10}$$

and L_1 loss is still the same as Equation 3.

3.4 Architecture

Our architecture, shown in Figure 3, is based on the one of the pix2pix method which uses a convolution-based U-Net [36] as its generator and a patch-wise discriminator. We add a proposed CSAM after every convolutional layer of the generator except the first and last ones. The conditional image is resized to specific size and concatenate to the previous feature maps as the input of every CSAM. CSAMs are able to access the information of the conditional image directly and model the long-range dependencies across images and feature maps. Also, we switch the patch-wise discriminator into the proposed multiple level patch discriminator to enable the discriminator network to capture both global and local information and therefore guide the generator to generator images with more structural layout. More details of the architecture are discussed blow.

Noise vector. Some past conditional GANs add a noise vector to the generator as input to avoid it producing a deterministic output. However, the pix2pix model has shown that the noise vector is just ignored by the generator network and hardly change the output samples. We observe the same phenomenon in our experiments and do not apply the noise vector in our model.

Spectral Normalization. Spectral normalization [30] is a recently proposed normalization technique, which restricts the spectral norm of each layer of the discriminator to constrain its Lipschitz constant. Spectral normalization is computationally efficient and require no extra hyper-parameter. It has shown that spectral normalization also benefit the training of generator by avoiding unusual gradients. We add spectral normalization to the discriminator and CSAMs in the generator.

4 EXPERIMENT

We propose the CSGANs framework, which translate images from one domain to another, being able to capture long range dependencies and reserve the global structures. To demonstrate the effectiveness of our framework, we have performed several experiments. In this section, we discuss

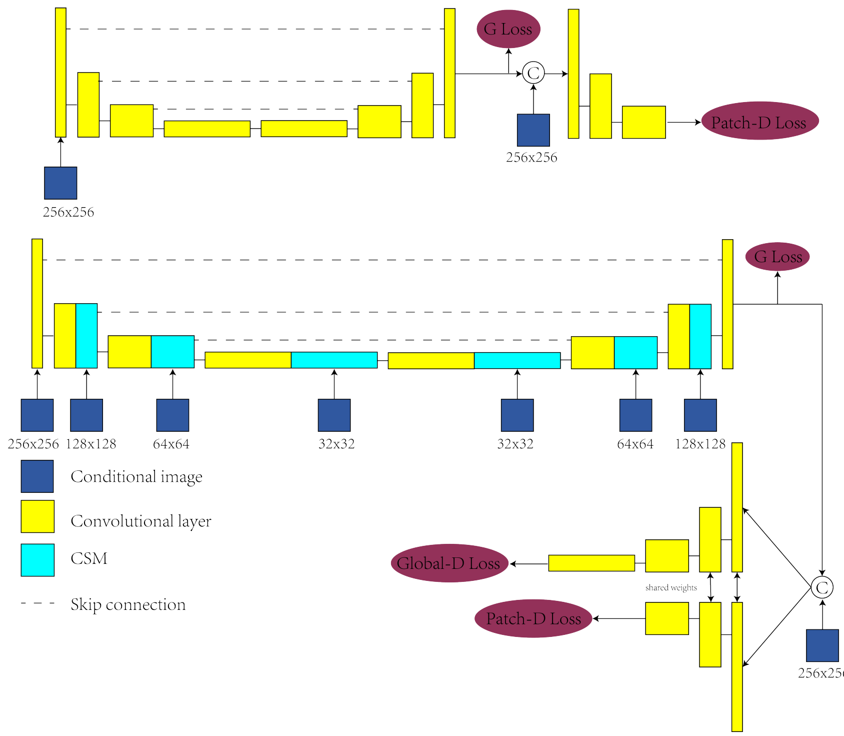

Fig. 3. Architecture

4.1 Implementation Details

In order to comparing with the pix2pix model, we basically follow its implementation details. We use minibatch SGD and Adam [20] optimizer with learning rate $lr = 0.002$ and momentum parameters $\beta_1 = 0.5, \beta_2 = 0.999$. We update one step for either of G and D alternatively. Batch normalization is used in convolutional layers of the generator. Batch size is set to 8.

4.2 Dataset

We evaluation our method with the task of translating edge maps to natural images, e.g. the target images are face images while the conditional images are the corresponding edge maps. The face images of the dataset we used are face images of CelebA dataset [**?**], which is a large-scale face attributes dataset with more than 200K celebrity images. Faces have well-defined structure of eyes, noses, mouths, and etc., and therefore the artifacts are visually sensitive for observers. This makes face images suitable for evaluating the proposed method. We utilize the cropped and aligned version of dataset with the size of every images being 218×178. In order to meat the original setting of the pix2pix method, we center-crop the images and resize the image to 256×256 in both experiments of the pix2pix model and the proposed model. The face attributes are attached in the dataset but not included in our experiments.

The edge maps we use generated in the pipeline similar to that used in pix2pix paper. Specifically, the edges are firstly extracted using a deep edge detector named holistically-nested edge detect (HED) [43]. We keep the values of each edge pixels calculated by HED in the edge maps. Each of these values is supposed to indicate the probability of being edge in the positions of pixel. And then several steps of post-processing are conducted to obtain simpler and clearer edge maps with fewer edge fragments, including thinning, short edge removal, and erosion. In addition, since the edge maps are very sparse, we add one more step to the process to decrease the sparsity of edge maps. We calculate an unsigned euclidean distance field for each edge map to obtain a dense representation. We note that similar idea of distance filed representations can be found in some recent works [14, 32**?**]. In Section **??**, we will prove the advantages of distance fields by experiments.

4.3 Evaluation Metrics

The evaluation of generative models is an open and complicated task, because a model with good performance with respect to one criterion need not imply good performances with respect to the other criteria [27, 38]. Traditional metrics, such as pixel-wise mean-squared error do not present the joint statistics of the synthesized samples and therefore is not able to evaluation the performance to a conditional generated model. Inception Score (IS) [**?**] is a widely-used criterion. However, IS has been pointed out to have serious limitations that it focuses more on the recognizability of the generated images rather than realism of details or intra-class diversity [38]. Moreover, IS is an evaluation metric for class-aware task which is not suitable for our experiments.

Since the goal of image-to-image translation is to generate from the conditional image an corresponding image visually plausible to human, we mainly compare the results between different models by perceptual user studies. Several related works have proposed similar perceptual experiments [4, 7, 23, 41**?**]. Following the similar procedure as described in [4], we conduct two different kinds of experiments: unlimited time user study and limited time user study. In addition, we use anther popular criterion, Fréchet Inception Distance (FID) [**?**], to prove the effectiveness of proposed method quantitatively. More details are explained below.

Unlimited Time User Study. We utilize perceptual user study experiments to compare the generated samples between different models. In every trial, we randomly select a conditional image from the testing dataset and generate two synthesized images from pix2pix and our model that are going to be compared with each other. These three images are displayed side by side, and the user is asked to pick one from the two synthesized images within unlimited time based on "which is more realistic and matches the conditional image better". The options offered to users are two of the synthesized images. No feedback is provided after every trial to avoid misguiding the user's perceptual judgment and preference. 00 users participate this experiments, and 00 trials are provided to every user.

Limited Time User Study. For this task, we evaluates how quickly the users can perceive the differences between images. In every comparison, we select three images corresponding to one randomly drawn edge maps (two generated by pix2pix and our model, and the ground truth). Similarly, two of these three images are displayed to the use with the edge map side by side for a short period of time. The user is asked to pick one of two displayed face images also based on "which is more realistic and matches the conditional image better". The duration is randomly selected between 1/8 seconds and 8 seconds.

Fréchet Inception Distance (FID). Fréchet Inception Distance (FID) [**?**] is a recently proposed and widely used evaluation metric for generative models, which is shown to be consistent with human perceptual evaluation in assessing the realism and variation of generated samples. FID uses an Inception network to extract features and calculates the Wasserstein-2 distance between features of the generated images and the real images. Models with lower FID values are supposed to model a synthetic distribution closer to the real distribution. We inference each model with the conditional

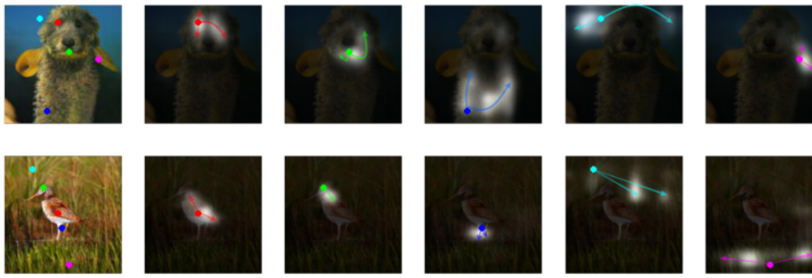

Fig. 4. This is a figure borrowed from SAGANs to show what kind of figure we are going to show here. The description will be added once the experiments are finished and the results are obtained.

images in the testing set to get the generated samples, and calculate the FID with respect to the target images in the testing set.

4.4 Comparison with pix2pix

In this section, experiments are conducted to compare the images generated by the pix2pix model and the proposed model. The comparisons are describe blow.

User Study. Two kinds of user studies are performed.The unlimited time user study is designed to evaluate the perceptual quality of the generated image, the results of which are shown in Table ??. We can observe that ....The limited time user studies are designed to evaluate how quickly the users cn perceive the differences between images. Figure **??** shows the results......

MS-SSIM and FID. xxxx

Table 1. Unlimited time user study.

	pix2pix	ours
user preference	0%	0%

Table 2. Evaluation metrics

Images	MS-SSIM	FID
Dataset	0	0
pix2pix	0	0
-Distance fields	0	0
-Spectral Normalization	0	0
-Global Discriminator	0	0
-Conditional Connection	0	0
Full model	0	0

4.5 Ablation study

We examine the importance of every part of our model by MS-SSIM and FID, shown in Table 2. Experiments are conducted by remove specific part from the full model and then generate images without this part. Specifically, we remove 1) the calculation of the unsigned distance fields of edge

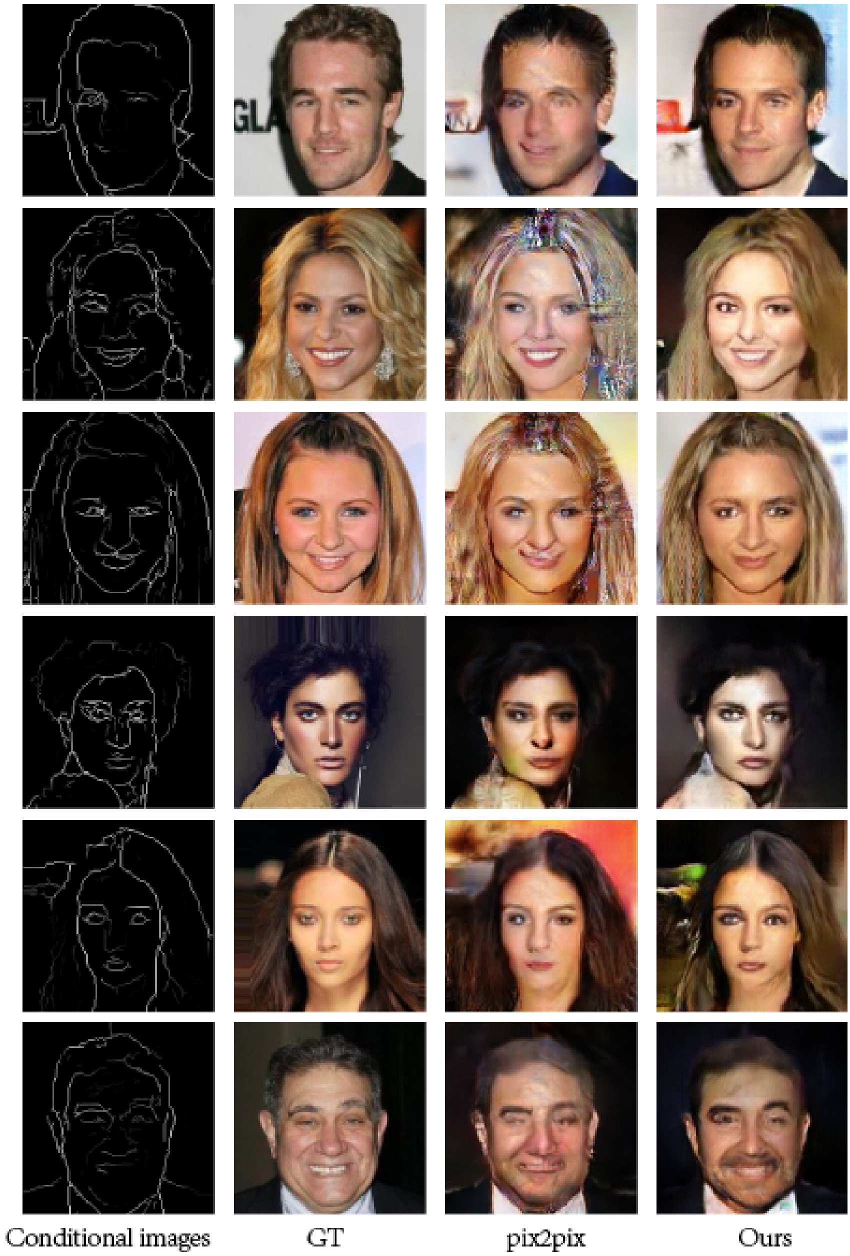

Fig. 5. results

maps before input to the generator and discriminator(-distance fields), 2) the spectral normalization in the CSAMs of generator and convolutional layers of discriminator (-spectral normalization), 3) the global discriminator in the multi-level discriminator (-global discriminator), and 4) the connection in each CSAM of resized images (-conditional connection). We observe that.

4.6 Self-Attention

We visualize the attention weight of CSAMs in the generator to find out how pixels of all locations in the images and feature maps are learned attend to one specific pixel. Figure 4 shows some examples.

5 CONCLUSION

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875* (2017).

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[3] David Berthelot, Thomas Schumm, and Luke Metz. 2017. BEGAN: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717* (2017).

[4] Qifeng Chen and Vladlen Koltun. 2017. Photographic image synthesis with cascaded refinement networks. In *IEEE International Conference on Computer Vision (ICCV)*, Vol. 1. 3.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2172–2180.

[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, Rob Fergus, et al. 2015. Deep generative image models using aï£ij laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*. 1486–1494.

[8] Alexey Dosovitskiy, Jost Tobias Springenberg, Maxim Tatarchenko, and Thomas Brox. 2017. Learning to generate chairs, tables and cars with convolutional networks. *IEEE transactions on pattern analysis and machine intelligence* 39, 4 (2017), 692–705.

[9] A. A. Efros and T. K. Leung. 1999. Texture synthesis by non-parametric sampling. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 2. 1033–1038 vol.2. https://doi.org/10.1109/ICCV.1999.790383

[10] W. T. Freeman, T. R. Jones, and E. C. Pasztor. 2002. Example-based super-resolution. *IEEE Computer Graphics and Applications* 22, 2 (Mar 2002), 56–65. https://doi.org/10.1109/38.988747

[11] L. A. Gatys, A. S. Ecker, and M. Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2414–2423. https://doi.org/10.1109/CVPR.2016.265

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14)*. MIT Press, Cambridge, MA, USA, 2672–2680. http://dl.acm.org/citation.cfm?id=2969033.2969125

[13] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. 2017. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. 5769–5779. http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans

[14] X. Han, Z. Li, H. Huang, E. Kalogerakis, and Y. 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. https://doi.org/10.1109/ICCV.2017.19

[15] James Hays and Alexei A. Efros. 2008. Scene Completion Using Millions of Photographs. *Commun. ACM* 51, 10 (Oct. 2008), 87–94. https://doi.org/10.1145/1400181.1400202

[16] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. 2017. Stacked generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2.

[17] P. Isola, J. Y. Zhu, T. Zhou, and A. 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

[18] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. 2016. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215* (2016).

[19] 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).

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[21] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[22] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. 2017. On convergence and stability of GANs. *arXiv preprint arXiv:1705.07215* (2017).

[23] C. Ledig, L. Theis, F. HuszÃąr, 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. https://doi.org/10.1109/CVPR.2017.19

[24] Chuan Li and Michael Wand. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*. Springer, 702–716.

[25] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised Image-to-Image Translation Networks. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 700–708. http://papers.nips.cc/paper/6672-unsupervised-image-to-image-translation-networks.pdf

[26] Ming-Yu Liu and Oncel Tuzel. 2016. Coupled Generative Adversarial Networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., USA, 469–477. http://dl.acm.org/citation.cfm?id=3157096.3157149

[27] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. 2017. Are gans created equal? a large-scale study. *arXiv preprint arXiv:1711.10337* (2017).

[28] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. 2017. Least Squares Generative Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2813–2821. https://doi.org/10.1109/ICCV.2017.304

[29] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[30] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957* (2018).

[31] Takeru Miyato and Masanori Koyama. 2018. cGANs with Projection Discriminator. In *International Conference on Learning Representations*. https://openreview.net/forum?id=ByS1VpgRZ

[32] D. T. Nguyen, B. S. Hua, M. K. Tran, Q. H. Pham, and S. K. Yeung. 2016. A Field Model for Repairing 3D Shapes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5676–5684. https://doi.org/10.1109/CVPR.2016.612

[33] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional Image Synthesis with Auxiliary Classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. PMLR, International Convention Centre, Sydney, Australia, 2642–2651. http://proceedings.mlr.press/v70/odena17a.html

[34] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, and Alexander Ku. 2018. Image Transformer. *CoRR* abs/1802.05751 (2018). arXiv:1802.05751 http://arxiv.org/abs/1802.05751

[35] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv preprint arXiv:1511.06434* (2015).

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[37] 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*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 2234–2242. http://papers.nips.cc/paper/6125-improved-techniques-for-training-gans.pdf

[38] Lucas Theis, Aäron van den Oord, and Matthias Bethge. 2015. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844* (2015).

[39] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. 2016. Conditional Image Generation with PixelCNN Decoders. *arXiv preprint arXiv:1606.05328* (2016).

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[41] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[42] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2017. Non-local neural networks. *arXiv preprint arXiv:1711.07971* 10 (2017).

[43] Saining "Xie and Zhuowen" Tu. 2015. Holistically-Nested Edge Detection. In *Proceedings of IEEE International Conference on Computer Vision*.

[44] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[45] 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

[46] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2018. Self-Attention Generative Adversarial Networks. *arXiv preprint arXiv:1411.1784* (2018).

[47] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. 2017. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. *CoRR* abs/1710.10916 (2017). arXiv:1710.10916 http://arxiv.org/abs/1710.10916

[48] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. StackGAN: Text to Photo-Realistic Image Synthesis With Stacked Generative Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 5907–5915.

[49] Junbo Zhao, Michael Mathieu, and Yann LeCun. 2016. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126* (2016).

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

[51] 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*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 465–476. http://papers.nips.cc/paper/6650-toward-multimodal-image-to-image-translation.pdf

Received February 2007; revised March 2009; accepted June 2009