Edge2photo

YUHANG LI, University of Science and Technology of China, China
XUEJIN CHEN, University of Science and Technology of China, China

This is an abstract example of this Latex template.«« Multifrequency media access control has been well understood in general wireless ad hoc networks, while in wireless sensor networks, researchers still focus on single frequency solutions. In wireless sensor networks, each device is typically equipped with a single radio transceiver and applications adopt much smaller packet sizes compared to those in general wireless ad hoc networks. Hence, the multifrequency MAC protocols proposed for general wireless ad hoc networks are not suitable for wireless sensor network applications, which we further demonstrate through our simulation experiments. In this article, we propose MMSN, which takes advantage of multifrequency availability while, at the same time, takes into consideration the restrictions of wireless sensor networks. Through extensive experiments, MMSN exhibits the prominent ability to utilize parallel transmissions among neighboring nodes.»»

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 and Xuejin Chen. 2010. Edge2photo. *ACM Trans. Web* 9, 4, Article 39 (March 2010), 6 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Realistic image synthesis has been a hot topic in computer vision and computer graphics for years. »>Some traditional image generation methods«< With the emergence of deep neural networks (DNN), several promising approaches for image synthesis have been proposed. Variational autoencoders (VAEs) [**?**] which maximize a variational lower bound on the log-likelihood of the training data. Autoregressive [**?**] networks directly model the conditional distribution over pixels. Though generating convincing samples, these models are costly to sample from.

Generative adversarial networks (GANs) [**?**] offer a new and promising mechanism to generate images, which take noises vectors as input and train two networks playing minmax game to guide the generated samples to be indistinguishable from the real ones. Conditional GANs are generalized versions of GANs in a conditional setting. Instead of noise vectors, cGANs generate images with different conditions, such as discrete class labels [**?**], texts [**?**], and xxx. Among these conditional image generation methods, image-to-image translation has drawn a lot of attention recently, which aims to apply a source image in one domain to generate a corresponding target image in another, reserving shared concepts, objects or scenes in these two images. Since [**?**]

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.

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

raised the first image-to-image model(pix2pix), there have been a wide variety of this approach [**?**]. However, image-to-image translation models based on convolutional neural networks may have troubles to generate some classes of realistic images, especially when these images have structural constrains, for example, generating faces from corresponding edge maps (edge-to-face) »»add a figure to explain««. The reasons behind this might be 1) that it is hard for the optimizer to discover parameter values that simulates the long range dependencies through several convolutional layers, since the convolutional operator has local receptive field [**?**], and 2) that the discriminator used in pix2pix models [**?**] 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 Self-attention [**? ? ? ?**], 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. For the second reason, we consider to establish the multiple patch-wise discriminators to capture information of different level. This can be achieved by modifying the loss of discriminator and not changing its architecture. We note that similar idea is raised by [**?**] 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. »» Discuss more details of the results and model «« Our contributions are summarized as follow:

i)

ii)

iii)

The rest of this article is organized as follow. Related works are presented in Section »» more detail «««

2 RELATED WORK

Our work is based on generative adversarial networks (GANs) [**?**] in a conditional setting. In this section, we present related research in GANs.

2.1 Generative Adversarial Networks (GANs)

Generative adversarial networks (GANs) [**?**] have obtained a great success in recent years. 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. Many recent works extent GANs to a wide range of applications, such as image generation [1, 42, 62], representation learning [45], image manipulation [64], object detection [33], and video applications [38, 51, 54].

2.2 Conditional Generative Adversarial Networks

Conditional GANs were firstly introduced by [**?**] 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 in the condition of discrete labels [**?**], text [**?**] and images. [**?**] 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. [**?**] generated photo-realistic images conditioned on text descriptions. Our work utilize GANs in a conditional setting to generate images from images.

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.

[**?**], called pix2pix, firstly introduced the concept of image-to-image translation, who treated one image in a paired image dataset as conditioned input and generate its corresponding image. Pix2pix applies skip connections [**?**] 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. »>Discuss the shortage of pix2pix«< [**?**] 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 [**?**], DiscoGAN [**?**] and DualGAN [**?**] 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. »»discuss the different between supervised and unsupervised methods«««<

2.4 Attention mechinism

»»» Not familiar yet. «««««<

3 METHOD

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

3.1 Preliminary

The pix2pix [?] is a image-to-image translation framework based on conditional GANs, which trains a generator network G and a discriminator network D. The generator G takes as input conditional images and outputs corresponding target images, while the discriminator D aims to distinguish real images from the synthesized ones. Formally, to train these two networks in a supervise manner, a set of pairs of corresponding images is required as training set (x_i, y_i), where x_i is a source image and y_i is a corresponding target image. These two networks play a minmax game:

$$\min_G \max_D L_{adv}(G, D) >>> L1loss <<< \tag{1}$$

to guide the generator to model the conditional distribution of real images given the source images, where the adversarial loss function is generally given by

$$E_{s\sim p_{data}(x)}[\log D(x,y)] + E_{(x,y)\sim p_{data}(x,y)}[\log(1 - D(x, G(x)))] \tag{2}$$

The generator of pix2pix is a convolution-based U-Net [**?**], the input of which is only applied to the first layer. The discriminator is patch-wise discriminator introduced by PatchGANs [**?**]. The conditional image is concatenated channel-wisely to the synthesized image or real image as the input of the discriminator. »» show the loss implementation in PatchGANs ««<

3.2 Conditional Self-Attention Module (CSM)

Since convolution operation focuses on local receptive field and have to capture the long rang dependencies across the entire image through several layers, which is computationally inefficiently. Inspired by [**?**] and [**?**], we introduce a conditional self-attention module (CSM) to the convolution-based pix2pix framework in order to capture the long rang dependencies of images and feature maps, as shown in Figure **??**

Given the image features from the previous hidden layer $a \in \mathcal{R}^{C\times M}$, we first resize the conditional image $x \in \mathcal{R}^{3\times N}$ to $x_a \in \mathcal{R}^{3\times M}$ and concatenate the resized conditional image to the image feature to get $[a, x_a]$ 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 focuses on important regions directly based on the conditional image. Then the conditioned features are mapped into three feature space by

$$f([a, x_a]) = W_f[a, x_a], \tag{3}$$

$$g([a, x_a]) = W_g[a, x_a], \tag{4}$$

$$h([a, x_a]) = W_h[a, x_a], \tag{5}$$

where $W_f, W_g \in \mathcal{R}^{\hat{C}\times(C+3)}$, $W_h \in \mathcal{R}^{(C+3)\times C}$ are trainable weights, which are implemented by 1×1 convolutions. Here, we use $\hat{C} = C/16$ in our experiments. Let $\beta_{j,i}$ be the indicator that indicates the extent to which the model attends to the i^{th} location when synthesizing the j^{th} region, which is calculated by

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

where $s_{ij} = f([a, x_a])^T g([a, x_a])$. Next, we use $\beta_{j,i}$ as the attention weights and compute the response $r = (r_1, r_2, ; r_M) \in \mathcal{R}^{\times M}$ at every position as a weighted sum of the features at all positions, where

$$o_j = \sum_{i=1}^{M} \beta_{j,i} h([a, x_a]). \tag{7}$$

As suggested in [**?**], 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

$$o_i = \gamma r_i + a_i, \tag{8}$$

where γ 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 rang dependencies by assign more weight to the non-local evidence progressively.

3.3 Multiple Level Patch Discriminator

The discriminator pix2pix uses is a patch-wise dicriminator [**?**], which distinguishes the real/synthesized images patch by patch with in a local receptive field much smaller than the size of the input images, and averages all responses to provide 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 structure of the conditional image are global information across image, 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 **??**.

3.4 Architecture Details

xxx

4 EXPERIMENT

We propose Conditional Self-attention Generative Adversarial Networks (CSGANs), 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 x sets of experiments. »»> A brief introduction of these experiments «««<

4.1 Implementation Details

4.2 Dataset

We evaluation our method with the task of translating edge maps to natural images, e.g. the target images are natural images while the conditional images are corresponding edge maps. The natural 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 are of well-defined structure of eyes, noses, mouths, and etc., therefore the artifacts are visually sensitive for observers. We utilize the cropped and aligned version of dataset with the size of every images being 218×178. 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 binary edge maps are firstly extracted using a deep edge detector named holistically-nested edge detect (HED) [**?**]. And then several steps of post-processing is 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 the 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 [**? ? ?**]. In section **??**, we will prove the advantages of the distance fields by experiments.

»» data split? ««

4.3 Evaluation Metrics

The evaluation of generative models is a open and complicated task, because the model with good performance with respect to one criterion need not imply good performances with respect to the other criteria [**? ?**]. 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. 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 [**?**]. However, IS is evaluation metrics

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. In addition, we use anther popular criterion, FrÃÍchet Inception Distance (FID) [**?**], to prove the effectiveness of proposed method quantitatively.

4.3.1 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 two methods 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 (and "equally well"). No feedback is provided after every trial to avoid misguiding the user's perceptual judgment and preference. About 00 users participate the experiments, and 00 trials are provided to every user.

4.3.2 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 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

4.5 Ablation study

5 CONCLUSION

Received February 2007; revised March 2009; accepted June 2009