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, University of xxxx, USA

Image-to-image translation, 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, has drawn a lot of interest recently. Models based on deep neural networks have shown to be powerful in generating visually plausible images. However, convolution-based models fail to synthesize some classes with structural constraints. One possible explain is that the convolutional operation has local receptive and require a stack of several convolutional layers to obtain a large receptive field. We propose conditional self-attention generative adversarial networks (CSAGANx) to capture the long-range dependencies and global structure information across images. This model 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 the model by two kinds of perceptual user studies and Fréchet Inception Distance (FID), and show that this model xxxx.

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), 10 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 non-parametric methods [**?**] oftern matching images patch-wisely with a database of existing images. With the emergence of deep neural networks (DNN), several promising DNN-bsed approaches for image synthesis have been proposed. Variational autoencoders (VAEs) [**?**], 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 [**?**] directly model the conditional distributions over pixels. Though generating convincing samples, these models are costly to sample from.

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, University of xxxx, xx Rd, xxx, xxx, xxxx, USA, SingBing.Kang@microsoft.com.

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

Generative adversarial networks (GANs) [**?**] 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 are generalized versions of GANs in a conditional setting. Instead of noise vectors, conditional GANs are able to be applied to generate images with a variety of conditions, such as discrete class labels [**?**], texts [**? ?**], and lower-resolution images [**? ?**]. Conditional GANs are trained in a supervised manner and shown to be able to model the conditional distributions with respect to the assigned conditions.

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 [**?**] raised the first image-to-image model (pix2pix), there have been many variants of this approach in both supervised and unsupervised manner [**? ? ? ?**]. 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, when generating faces from corresponding edge maps (edge-to-face), convolutional-based pix2pix model sometimes fails to xxxxx »»add a figure to explain««. The reasons behind this might be two-folded. 1) 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 [**?**]. 2) The discriminator used in the pix2pix model [**?**] 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 [**? ? ? ?**], 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 [**?**] 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)

iii)

The rest of this article is organized as follow. Related works are presented in Section **??**. The method we proposed is introduced in Section **??**. We show the experiments to prove the effectiveness of method in Section **??**

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) [**?**] 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 [**? ?**], 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) [**?**] first introduced a convolutional architecture which led to improved visual quality. [**?**] proposed an approach to train discriminator in a semi-supervised fashion, granting the discriminatorâĂŹs internal representations knowledge of the class structure of (some fraction of) the training data it is presented. Energy based GANs (EBGANs) [**?**] 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 [**? ?**] to benefit both the stability and mode coverage. Several other works [**? ? ?**] also make progress in stabilizing the training and increasing the diversity of the results of GANs. »»SAGAN??««

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 [**?**] 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. [**? ?**] 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. [**?**] 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 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. Recently proposed [**?**] 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 [**?**] 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 [**?**] 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. [**? ?**] used coarse-to-fine refinement frameworks to synthesis photographic images from semantic label maps. [**?**] studied the generating images of outdoor scenes from semantic label maps coupled with attributes. [**?**] 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 [**?**]. [**?**] 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.

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 [**? ? ? ?**]. Self-attention [**? ?**] has been shown to be powerful in a variety of tasks. [**?**] applied self-attention to machine translation models, and demonstrated the plausible effectiveness of self-attention mechanism. [**?**] studied on combining the self-attention mechanism and autoregressive models, and proposed an image transformer model in image generation. Inspired by non-local operation in computer vision, [**?**] utilize self-attention mechanism as a non-local operation to model long-range spatial-temporal dependencies for video processing. [**?**] 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.

In particular, self-attention [4, 20], also called intra-attention, calculates the response at a position in a sequence by attending to all positions within the same sequence. Vaswani et al. [32] demonstrated that machine translation models could achieve state-of-the-art results by solely using a

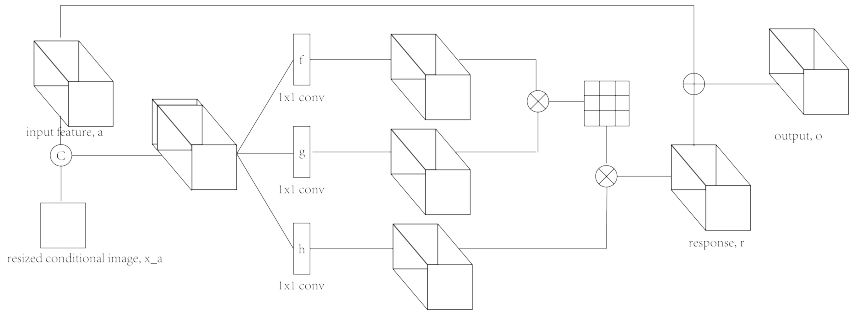

Fig. 1. CSM

self-attention model. Parmar et al. [21] proposed an Image Transformer model to add self-attention into an autoregressive model for image generation. Wang et al. [33] formalized self-attention as a non-local operation to model the spatial-temporal dependencies in video sequences. In spite of this progress, self-attention has not yet been explored in the context of GANs.

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 across image. 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 model [?] is an 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:

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

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

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

and the L_1 loss is given by

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

The generator of pix2pix is a convolution-based U-Net [**?**]. The condition image, the input of the generator, 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.

3.2 Conditional Self-Attention Module (CSM)

Since focusing on local receptive field, convolution operation has to establish a large receptive field through several layers, which is computationally inefficiently. It is possibly difficult for the optimizer to discover parameter values to model the relationship across the image. Inspired by [**?**] and [**?**], we introduce a conditional self-attention module (CSM) to the convolution-based pix2pix framework in order to capturing the long-range dependencies of images and feature maps, as shown in Figure ??

Given the image features from the previous hidden layer $a \in \mathbb{R}^{C\times M}$, we first resize the conditional image $x \in \mathbb{R}^{3\times N}$ to $x_a \in \mathbb{R}^{3\times M}$ 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

$$f([a, x_a]) = W_f[a, x_a], \tag{4}$$

$$g([a, x_a]) = W_g[a, x_a], \tag{5}$$

$$h([a, x_a]) = W_h[a, x_a], \tag{6}$$

where $W_f, W_g \in \mathbb{R}^{\hat{C}\times(C+3)}$, $W_h \in \mathbb{R}^{(C+3)\times C}$ 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

$$\beta_{j,i} = \frac{exp(s_{ij})}{\sum_{i=1}^{M} exp(s_{ij})} \tag{7}$$

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 $r = (r_1, r_2, \dot{;} r_M) \in \mathbb{R}^{\times M}$ at every position as a weighted sum of the features at all positions, where

$$o_j = \sum_{i=1}^{M} \beta_{j,i} h([a, x_a]). \tag{8}$$

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

$$o_i = \gamma r_i + a_i, \tag{9}$$

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 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. 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 structure of every conditional image 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

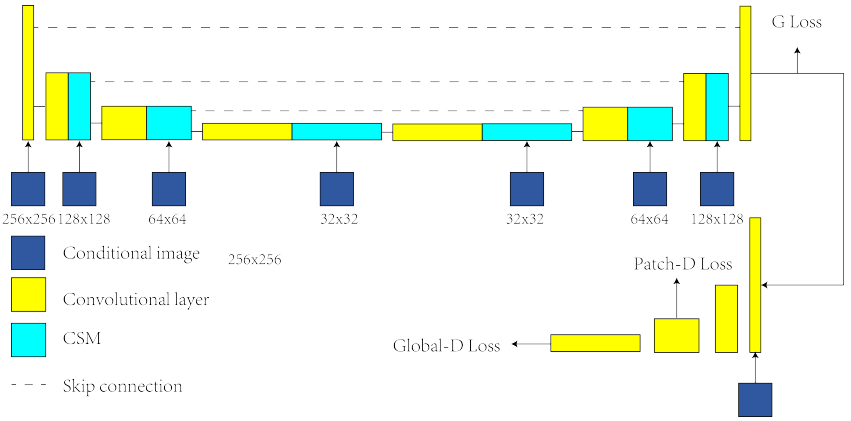

Fig. 2. Architecure

in first few layers since the lower features of these discriminators should be the same, as shown in Figure **??**.

3.4 Architecture

Our architecture is based on the architecture of the pix2pix method which use a convolution-based U-Net [**?**] as its generator and a patch-wise discriminator. We add the proposed CSM after every convolutional layers to the generator except the first and last ones. CSMs 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. Figure **??** shows the

3.5 Other Techniques

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 showed that the noise vector is just ignored by the generator network and not able to change the output samples. We observe the same phenomenon and do not apply the noise vector in our model.

Spectral Normalization. Spectral normalization [**?**] is recently proposed normalization techniques, 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 is shown that spectral normalization also benefit the training of generator by avoiding unusual gradients. We add spectral normalization to the discriminator and CSMs in the generator.

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 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 [**? ?**]. 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 [**?**]. 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 [**? ? ? ? ?**]. Following the similar procedure as described in [**?**], 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. »»> Discuss the results ««««

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. »»» Discuss the results ««««

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

Table 1. Evaluation metrics

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

4.5 Ablation study

5 CONCLUSION

Received February 2007; revised March 2009; accepted June 2009

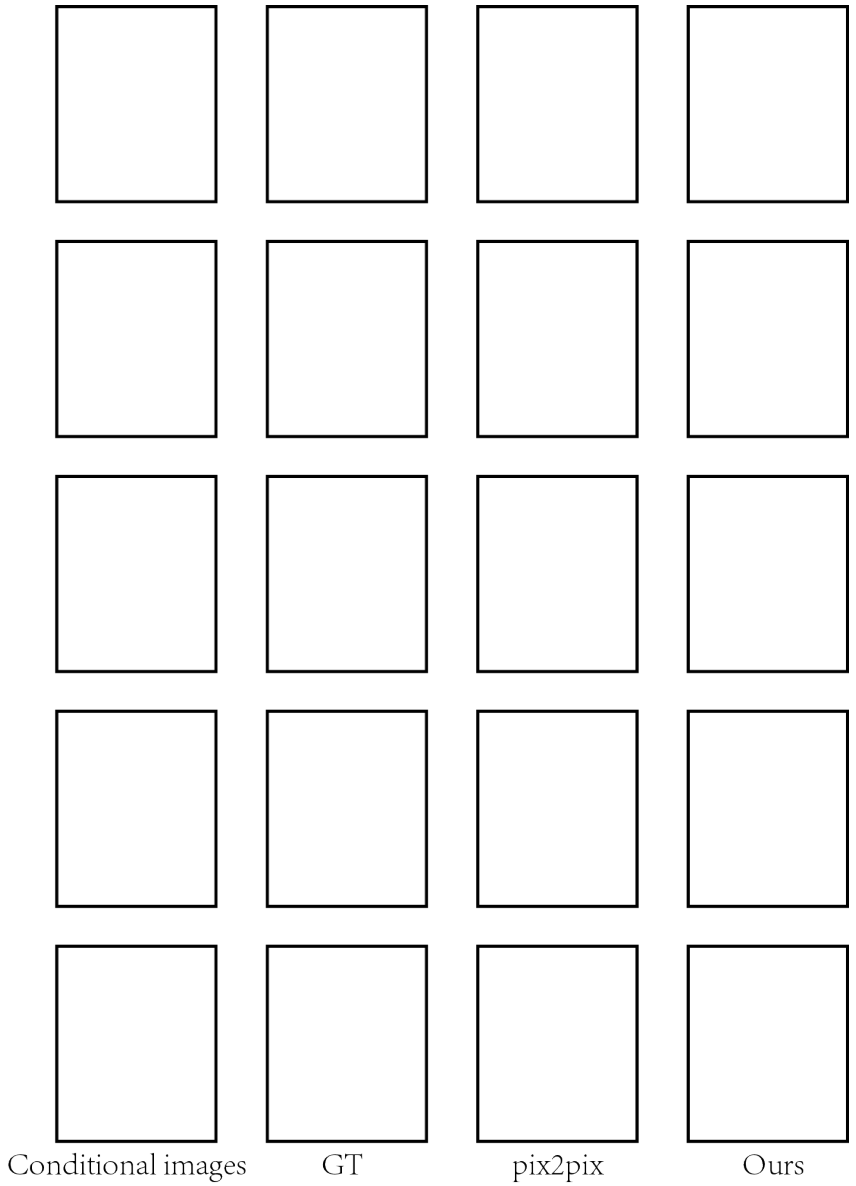

Fig. 3. results