Received August 1, 2019, accepted August 28, 2019, date of publication September 5, 2019, date of current version September 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939654

Consistent Embedded GAN for Image-to-Image Translation

FENG XIONG, QIANQIAN WANG, AND QUANXUE GAO
State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China

Corresponding authors: Qianqian Wang (qianqian174@foxmail.com) and Quanxue Gao (qxgao@xidian.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61773302 and Grant 61906142, in part by the Innovation Fund of Xidian University under Grant 10221150004, and in part by the Fundamental Research Funds for the Central Universities.

ABSTRACT Generative Adversarial Networks (GANs) have achieved remarkable progress in image-to-image translation tasks. However, these methods have the common problem that lacking the ability to generate both perceptually realistic and diverse images in the target domain. To tackle the problem, in this paper, we propose a novel model named Consistent Embedded Generative Adversarial Networks (CEGAN) for the image-to-image translation task. It aims to learn conditional generation models for generating perceptually realistic outputs and capture the full distribution of potential multiple modes of results by enforcing tight connections in both the real image space and latent space. To achieve realism, unlike existing GANs models that their discriminators attempt to differentiate between real images from the dataset and fake samples produced by the generator, the discriminator in our model distinguishes the real images and fake images in the latent space to alleviate the impact of the redundancy and noise in generated images. On the other hand, we learn a low-dimensional latent code that is distilled from the possible multiple distribution in the latent space to achieve diversity. By this way, our model avoids the problem of mode collapse and produces more diverse and realistic results. Extensive experimental results demonstrate the superiority of the proposed method.

INDEX TERMS Image-to-image translation, GAN, latent space.

I. INTRODUCTION

Nowadays, image-to-image translation tasks have attracted much attention in many computer vision articles due to its extraordinary performance [1]–[3]. It aims to learn a mapping that can convert an image from a source domain to a target domain, while preserving the main presentations of the input images. For instance, networks have been used to translate real-world scenes into cartoon images [4], add color to grayscale images [1], [5], [6], and fill missing image regions [3], [7], [8].

The goal of generative adversarial networks GANs [9] is to generate samples that can confuse the discriminator to achieve the purpose of falsehood to be dressed up as truth. It has achieved impressive success in images editing [7], super resolution [10], representation learning [11], and image generation [12], [13]. Particularly, the GANs are extensively studied in image-to-image translations. [14]–[16] tackle image-to-image translation by GANs where it used to ensure the generated images belonging to the target domain and improve image qualities by minimizing reconstruction loss.

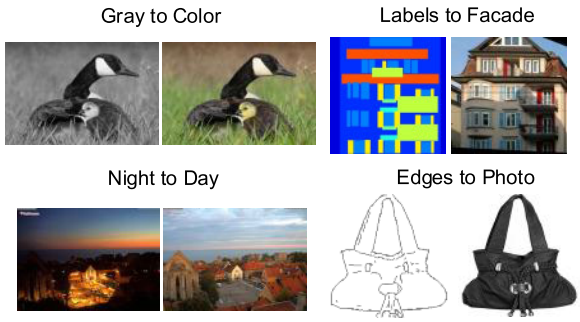

FIGURE 1. Many image generation problems involves translating an input image into a corresponding output image with GANs. Here, we show the results of image gray to color, labels to facade, night to day and edges to photo.

In recent years, several researchers have devoted to image-to-image translation with GANs field (see Figure 1) and excellent algorithms have emerged. The detailed methods

The associate editor coordinating the review of this manuscript and approving it for publication was Chang-Tsun Li.

are mainly divided into two categories. The first issue is that lots of articles have focused on learning one-to-one mapping from input to output. For example, pix2pix [3] actively learns a mapping from given input to output image, with a reconstruction loss to produce a similar output to the known paired ground truth image and a adversarial loss to encourage realism. It has previously been shown to produce good-quality results for a variety of image-to-image translation tasks. Similar ideas have been applied to various tasks such as generating photographs from sketches, attribute and semantic layouts [17]. However, these algorithms learn the mapping with paired training examples that are not easily got in real-world. Zhu *et al.* [16] tackles the unpaired setting with cycle consistency loss to retain the main presentations of image. CoGAN [18] and crossmodal scene networks [19] use a weight-sharing strategy to learn a common representation across domains. However, all above methods have focused on generating a single result conditioned on the input and these techniques usually assume a deterministic or unimodal mapping. As a result, they fail to capture the full distribution of possible outputs. Even if the model is made stochastic by injecting noise, the network usually learns to ignore it [20].

On the other hand, one way to help address the first issue is to leverage additional information from other modalities, so many interesting problems are more naturally thought of as a probabilistic one-to-many mapping. Lin *et al.* [21] is the first image-to-image work that decomposes image into domain-independent features and domain-specific features, and produces diverse translated images. In recent years, one of the most outstanding representatives of one-to-many algorithms is BicycleGAN [2], it learns a distribution of possible outputs and use it as cGANs model setting. Random sampling the learned ambiguous mapping distribution to produce the diverse outputs during the test time. The discriminator of BicycleGAN attempts to differentiate between real images from the dataset and fake samples produced by the generator, which is to ensure the generated images more similar to the groundtruth. Huang *et al.* [22] assumes that the image representation can be decomposed into a content code that is domain-invariant, and a style code that captures domain-specific properties. It recombines its content code with a random style code sampled from the style space of the target domain. These methods all realized the goal of producing multimodal outputs for a fixed input.

As we known, there are two main goals of the multimodal image generation problem: producing results which are perceptually realistic and diverse, while remaining faithful to the input. But this multimodal mapping in existing approaches leads to the common problem of mode collapse [9], where the generator learns to generate only a small number of unique outputs. Furthermore, although the above mentioned algorithms could get multimodal outputs conditioned on the same input, the quality of the generated images is unsatisfactory. The reason may be that in real-world application, the generated images by existing GANs always contain noise and redundancy because they directly discriminate the difference between generated image and real image in original space. However, the original images are usually high-dimensional images that may well contain redundant features, noise and outlying entries(In this article, we hold that redundant features, noise and outlying entries are not only image occlusion or destruction of pixel, but also blurred or unrealistic images. e.g., marked by red rectangle in Figure 2). The discriminator which identifies in original image space directly mainly considers the error relationship between the generated samples and the noisy samples. Low-quality generated images would seriously affect the function of discriminator. As a result, the final performance would be deteriorated greatly.

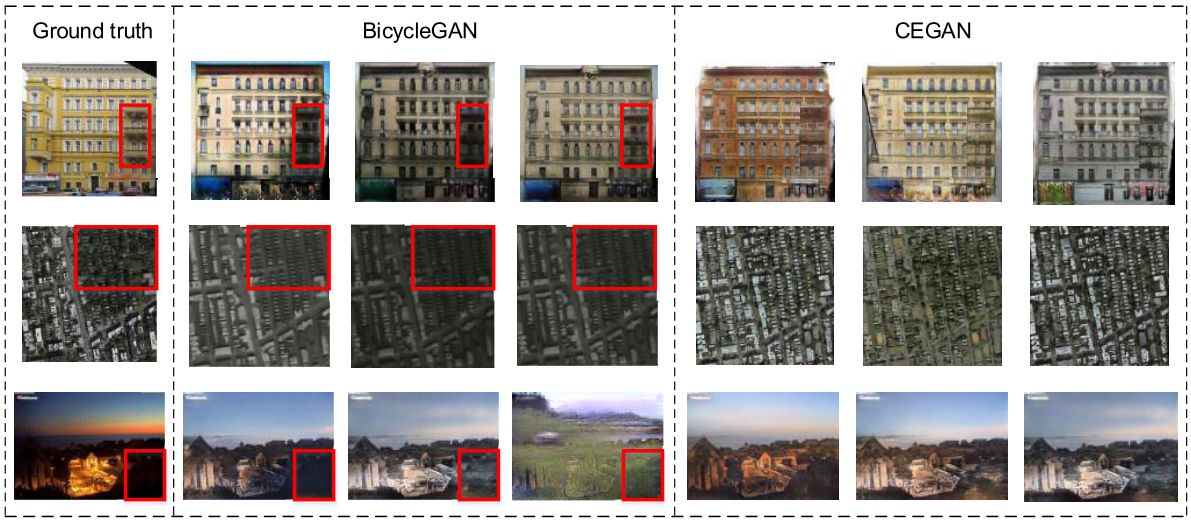

FIGURE 2. The explanation of noise, redundancy and outlying entries(marked by red rectangle) in original image space and generated images. Comparison of randomly generated samples from BicycleGAN and our CEGAN on Labels-facades, Map-satellite and Outdoor night-day image datasets.

In this work, our focus is to learn conditional generation models for generating perceptually realistic outputs and model a distribution of potential multiple modes of results by enforcing tight connections in both real image space and latent space. To ensure our produced images become real enough, unlike existing GANs models that the discriminator attempts to differentiate between real images from the dataset and fake samples produced by the generator, the discriminator in our proposed Consistent Embedded GAN(CEGAN) distinguishes the real images and fake samples in the latent space. We learn a low-dimensional latent code that is distilled from the possible multiple modes in the latent space to solve the problem of mode collapse and produces more diverse results. We choose GAN model and latent space learning due to the following considerations: (1) GAN can ensure that the generated images well mimic the natural images in the target domain. (2) Latent space learning model can help to alleviate the impact of redundance and noise in the generated images and produce more perceptually realistic and diverse outputs. The main contributions of our CEGAN model are summarized as follows:

- We propose a novel image-to-image translation model by combining GAN and latent space learning, our model can generate both realistic and diverse images.
- The discriminator in CEGAN distinguishes the real images and fake samples in the latent space instead of the real image space.
- We learn a mapping between real image space and latent space. Random sampling the ambiguity mapping to express multiple modes in the output.
- Extensive experimental results verify that our model outperforms the state-of-the-art image-to-image translation model.

II. RELATED WORKS

More and more GANs have been used in image-to-image translation application. We employ GANs to align the distribution of latent vector for generated images and real images in this paper. At the first, we will introduce GANs, then image-to-image translation, and multimodal encoding diversity.

A. GENERATIVE ADVERSARIAL NETWORKS (GANs)

The GANs [9] have achieved remarkable achievements in image generation field. During the training process, a generator is trained to fool a discriminator which in turn tries to distinguish between generated samples and real samples. In order to improve the quality of generated images, variant GANs have been proposed, such as better training objectives [23]–[25], combination with auto-encoders [26], [27] and multi-stage generation [28]–[30].

B. IMAGE-TO-IMAGE TRANSLATION

The task of image-to-image translation is to convert an image from a source domain to a target domain while preserving its certain properties. Mirza and Osindero [31] propose the first unified model for image-to-image translation based on conditional GANs, which has been successfully applied to many applications. For example, Wang *et al.* [32] apply it in generating high-resolution images. Condit ional GANs [31] demonstrated that both generator and discriminator are conditioned on some extra information to generate the output we expected. Similarly, conditional VAE [33] aims to translate the source domain to the target domain by adding random noise to the given image. Potentially, all of the methods defined above could be easily conditioned and have shown promise, while image-to-image conditional GANs have lead to a substantial boost in the quality of the results [21], such as pixel values [34], semantic features [35], class labels [36], or pairwise sample distances [37]. In addition, Liu *et al.* [38] propose the UNIT framework, which assumes a shared latent space such that corresponding images in two domains are mapped to the same latent code.

C. MULTIMODAL ENCODING DIVERSITY

As we know, a significant limitation of most existing image-to-image translation methods is insufficient diversity of the generated images. To tackle this problem, some scholars proposed a multi-output GAN model which can generating multiple different images from one input image [39], [40]. The most common way to generate multimodel outputs is to encode the equivocal samples in latent space which is conditioned on some mode-related context and input image. However, these methods can only generate a discrete numerical outputs. Zhu *et al.* [2] propose a BicycleGAN which can generate continuous and multimodal outputs. Although BicycleGAN has successfully generated multimodal outputs, the reality and diversity of generated image is still far from we expected. The main reason maybe that the discriminator discriminates the error relationship of distribution between the generated data and the noise-containing data. Latent space learning can effectively avoid the influence of redundant features and noise in generated images [41]. Thus, it motivates us to embed the multimodal encoding vector to the latent space. By learning a mapping between real image space and latent space, we random sample the ambiguity mapping to express multiple modes in the output.

III. CONSISTENT EMBEDDED GAN

In this section, we will present the detail of our model architecture.

A. MOTIVATION

Most existing GAN frameworks usually consist of two Convolutional Neural Networks (CNNs). One is the generator G which is trained to produce output that confuses the discriminator. The other is the discriminator D which classifies whether the image is from the real target manifold or synthetic. However, in real-world application, the original images are usually high-dimensional images that may well contain redundant features or noise. The traditional GAN structure mainly considers the error relationship between the generated image and the noisy image, which leads to noise

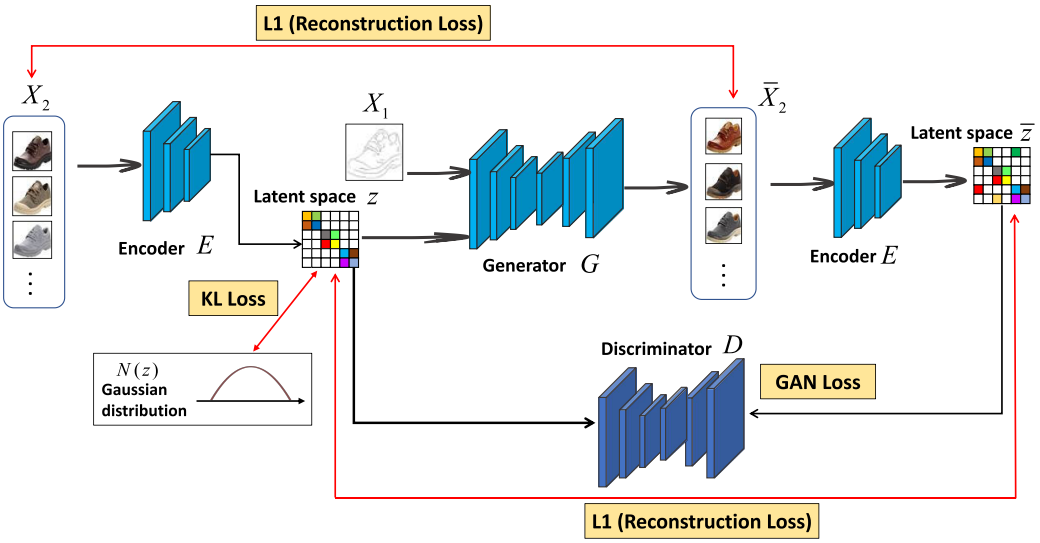

FIGURE 3. Modal overview. Our model CEGAN consists of a generator G, a discriminator D and a autoencoder E (the two E in the picture are essentially the same). First, ground truth X_2 is encoded into the latent space and obtain a latent code z. X_1 combines the latent code z as the input of G. The generator then attempts to map the input image X_1 along with the latent code z back into original image X_2. We encode the generated sample output $\overline{X_2}$ to be latent code $\overline{z}$ by E, that is to reduce the impact of generated noise and redundancy. We try to reconstruct both the ground truth's latent code and image to realize the purpose of generating realistic and diverse images in the target domain.

and redundancy in the generated images. As a result, the quality of generated images are unsatisfactory. While through latent space learning, we can transform high-dimensional image space into low-dimensional potential space and the obtained latent code can capture the main semantic information. The low-dimensional latent code adversarial learning is more conducive to the network training. So in the following section, we will introduce our method to alleviate such a challenging problem in detail.

B. CONSISTENT EMBEDDED GAN NETWORKS

Figure 3 shows an overview of our model. In the training progress, let $x_1 \in X_1$ and $x_2 \in X_2$ be the images from two different image domains, which are a dataset of paired images and are representative of a joint distribution $p(x_1, x_2)$. We should learn a multi-modal mapping between two image domains, for example, X_1 and X_2 represent edges and ground truth photographs respectively, and we want to generate a set of photographs about X_2 which have different colors and textures according to the edges of X_1. To achieve this, we train G to translate an input images x_1 into an output images x_2 conditioned on the target domain images' latent vector. It is important to note that there could be multiple plausible paired images x_2 which would correspond to an input images x_1 but the training dataset usually contains only one such pair. However, given a new image $\overline{x_1} \in X_1$ during test time, our model CEGAN would be able to generate a diverse set of output $\overline{x_2} \in X_2$, corresponding to different modes in the distribution $p(x_2 | x_1)$.

We would like to learn the mapping that could sample the output $\overline{x_2}$ from true conditional distribution given $\overline{x_1}$, and produce results which are both diversity and realism. In order to achieve diversity, we learn a low-dimensional latent code z that encapsulates the ambiguous aspects of the output mode which are not present in the input image. According to the latent code z, we could get different styles with the same input. We then learn a deterministic mapping $F : (x_1, z) \rightarrow x_2$. To enable stochastic sampling, we desire the latent code to be drawn from some prior distribution $p(z) \triangleq E(x_2)$. On the other hand to achieve realism, unlike the existing GANs framework that the discriminator D attempts to differentiate between real samples and generated samples, the discriminator D in CEGAN distinguish the real images and fake samples in the latent space. As we all know that during training process, the generate images always contain redundancy features, noise and outlying entries, which would lead to unreliable and inaccurate results. By latent space learning, we encode this images to the low-dimensional latent code z to alleviate such a challenging problem. Then discriminator D classifies whether the latent code is from the real target manifold or synthetic. Furthermore, to further ensure the quality of the generated images, we use a standard Gaussian distribution $\mathcal{N}(0, I)$ to constraint the latent distribution.

C. LOSS FUNCTIONS

Our loss functions comprise an adversarial loss, two reconstruction losses and a KL loss. GAN adversarial loss drives the generator network to match the distribution of translated images to the desired domain image distribution. Reconstruction loss and KL loss ensure the generated images similar to the known paired ground truth image. We use a simple additive form for the loss function:

$$L^* = \arg\min_{G,E}\max_{D} l_{GAN}(G, E, D) + \lambda_{image} l^{image}_{recon}(G, E) + \lambda_{latent} l^{latent}_{recon}(G, E) + \lambda_{KL} l_{KL}(E), \quad (1)$$

where the hyper-parameters λ_{image}, λ_{latent} and λ_{KL} control the importance of each term.

1) ADVERSARIAL LOSS

We employ conditional GANs to realize image-to-image translation task. In other words, images generated by our model should be indistinguishable from real images in the target domain. What's more, in order to better reflect the role of discriminator D, our D distinguish the real images and fake samples in the latent space that can decrease the impact of redundancy features and noise in the generated images. We illustrate the adversarial loss below:

$$l_{GAN} = \mathbb{E}_{z\sim p(z)}[\log(D(z))] + \mathbb{E}_{x_1\sim p(x_1),z\sim p(z)}[\log(1 - D(E(G(x_1, z))))]. \quad (2)$$

2) RECONSTRUCTION LOSS

To encourage the output of the generator to match the input as well as stabilize the GANs training, we use an l_1 loss to restrain the output image and the ground truth. Furthermore, to ensure the generated images' realism and diversity, we also reconstruct the output's latent vertor and the ground truth image's latent vector.

$$l_{recon}^{image} = \mathbb{E}_{x_1,x_2\sim p(x_1,x_2),z\sim p(z)}\|x_2 - G(x_1, z)\|_1, \quad (3)$$

$$l_{recon}^{latent} = \mathbb{E}_{x_1\sim p(x_1),z\sim p(z)}\|z - E(G(x_1, z))\|_1. \quad (4)$$

3) KL LOSS

To maximize the effectiveness of latent distribution and sample z at inference, we restrain it with a Gaussian assumption and encourage to restructure with KL dispersion.

$$l_{KL} = \mathbb{E}_{x_2\sim p(x_2)}[\mathcal{D}_{KL}(E(x_2) \parallel \mathcal{N}(0, I))]. \quad (5)$$

IV. IMPLEMENTATION

We implemented our CEGAN framework in PyTorch [42].

A. NETWORK CONFIGURATION

CEGAN is constructed with identical network architecture for G, D and E. For generator, it is configured with equal number of downsampling and upsampling layers. In addition, we configure the generator with symmetric skip connections between downsampling and upsampling layers as in BicycleGAN, making it a U-Net [43]. Such a design has been shown to produce strong results in the unimodal image prediction setting since it enables low-level information to be shared between input and output pairs. Without the skip layers, information from all levels has to pass through the bottleneck, typically causing significant loss of high-frequency information. For discriminator, we employ three fully connected layers, which aims to predict the real or fake latent code rather than images or overlapping image patches. Such a configuration is effective in capturing low-dimensional latent code distribution and it fulfills our needs well. For the encoder, it includes several strided convolutional layers to downsample the input, and a few residual blocks to further process it, followed by a global average pooling layer and a fully connected layer.

To stabilize our model training procedure, we build our model on the Least Squares GANs (LSGANs) variant [24], which uses a least-squares objective instead of a cross entropy loss. LSGANs produces high quality results with stable training process.

B. HYPERPARAMETERS

We use Adam optimizer [44] to train our networks with an initial learning rate of 0.0002. The learning rate is decreased by half every 10000 iterations. In all experiments, we use a batch size of 1 and set loss weights to $\lambda_{image} = 10$, $\lambda_{latent} = 1$ and $\lambda_{KL} = 0.1$. We choose the dimension of the latent code to be 8 across all datasets. Random mirroring is applied during training.

C. INJECTING THE LATENT CODE TO GENERATOR

To realize the diversity of outputs, we encode the possible multiple outputs in the latent space and combine the latent code with the given image as the input of the generator. By learning a mapping between real image space and latent space, we random sample the ambiguity mapping to express multiple modes. So how to propagate the information encoded by latent code to the image generation process is critical to our applications. There are two common solutions in existing methods. The most simply strategy is to extend a Z-dimensional latent code to an $H \times W \times Z$ spatial tensor and concatenate it with the $H \times W \times 3$ input image. Alternatively, the other method is to add the latent code to each intermediate layer of the network G. In this paper, we chose the former because the experiment results are not much different but the first strategy is easy to implement. The overview framework of injecting the latent code is shown in Figure 4.

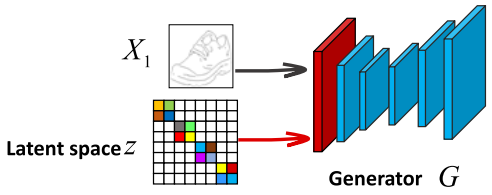

FIGURE 4. Injecting the latent code to generator. We inject the latent code z into the first input layer in the encoder.

V. EXPERIMENTS

A. DATASETS

Edges—Shoes/Handbags: We use the datasets provided by [45] and [46], which contain images of shoes and handbags with binary edge generated by the HED edges detector [47]. All the images are revised to 256×256 for our model training.

Map—Satellite: The dataset is provided by [3]. It collects the maps from Google Maps. Similarly, the images are downsampled such that the shortest side of each image is 256 pixels.

Outdoor Night—Day Images: The dataset comes from [48]. It contains 8571 images and each of which is 256 × 256.

Labels—Facades: We use the datasets provided by [49]. It is from architectural labels to photo and all the training images are revised to 256 × 256, trained on CMP Facades [50].

B. EVALUATION METRICS

AMT Perceptual Study: In order to compare the faithfulness and realism of translation outputs generated by different methods, similar to Wang *et al.* [32], we take human perceptual study on Amazon Mechanical Turk (AMT): the Turkers are presented with a series of trials that pitted an input image and six translation outputs from different methods. During each trial, the images appeared 2 seconds and Turkers are given unlimited time to respond which translation output looks more accurate. The appeared images are 256 × 256 resolution. We prepare 200 questions and select 100 people for our test. Each of Turker randomly selects 50 questions to answer.

LPIPS Distance: LPIPS distance [51] is one of the universal indicators for measuring image translation diversity. In our experiment, we compute the average LPIPS distance between pairs of randomly-sampled translation outputs from the same input as in Zhu *et al.* [2]. LPIPS is given by a weighted L2 distance between deep features of images. The ImageNet-pretrained AlexNet [52] extracts image feature fast with best performance, so we choose it as our experiment deep feature extractor. LPIPS distance has been demonstrated to correlate well with human perceptual similarity [51]. For each algorithm, we select 50 input images and every per input randomly generate 20 outputs. We choose the average LPIPS distance and standard deviation with the total 1000 pairs as the final result.

FID Score: FID score [53] is a measure of similarity between two datasets of images. It is calculated by computing the Fŕęchet distance between two Gaussians fitted to feature representations of the Inception network. It was shown to correlate well with human judgement of visual quality and was most often used to evaluate the quality of samples of Generative Adversarial Networks. In this paper, we choose the 768 pre-aux classifier features of the Inception network to calculate the FID distance.

C. BASELINES

cVAE-GAN: This method combines a variational autoencoder with a generative adversarial network to translate the images from source domain to target domain. It models an image as a composition of label and latent attributes in a probabilistic model. By varying the fine-grained category label fed into the resulting generative model to realize image style translation task.

cLR-GAN: This is another approach to capture image mode in latent space. It starts with a randomly sampled latent encoding, the conditional generator should result into an output which when given itself as input to the encoder should result back into the same latent code, enforcing self-consistency. This method is to explicitly model the inverse mapping and could be seen as a conditional formulation of the ''latent regressor'' model.

BicycleGAN: The method realizes bidirectional mapping by combining cVAE-GAN and cLR-GAN. It primely learns the relationship between the latent space and the real image space. It is the best existing image-to-image translation model we are aware of that can generate relative reality and multimodal outputs.

D. RESULTS

First, we qualitatively compare CEGAN with the three baselines above respectively. We test the performance of these methods with the same input and under the same training process. Figure 5 presents their translation results on Edges-shoes dataset. From Figure 5, we obtain the following observations:

- cLR-GAN indeed produces a relatively realistic output but the result shows less variation. Particularly, it sometimes suffers from mode collapse, e.g., the 2-th row and 6-th row blue rectangle outputs.
- On the other hand, cVAE-GAN adds variation to the output, as the latent space is encouraged to encode information about ground truth. However, the diversity of generated images comes at the cost of output's quality(e.g., marked by red rectangle).
- BicycleGAN sometimes achieves relatively more realistic and diverse transition than cVAE-GAN and cLR-GAN. However, the image quality of BicycleGAN is still unsatisfactory. This reason may be that it mainly considers the error relationship between generated data and noisy data.
- Images produced by our CEGAN model are both diverse and realistic. It has the best performance among all the compared methods, since CEGAN transforms high-dimensional image space into low-dimensional potential space, distinguishing the real and fake samples in the latent space. It enforces tight connections in both the real image space and latent space.

More results of CEGAN are shown on Figure 6. Quantitative evaluations confirm the qualitative observations above. We can obtain similar conclusions on Edges—handbags, Map—satellite and Night—day datasets which are shown in Table 1. Under the same training process on Edges-shoes dataset, we respectively use AMT perceptual scores to measure quality and LPIPS distance to evaluate diversity. As presented in the Table 1, cVAE-GAN and cLR-GAN get lower AMT perceptual scores than BicycleGAN and CEGAN. Moreover, cLR-GAN produces very little diversity according to LPIPS distance. Especially on Outdoor night—day image dataset, its LPIPS distance is around 12% below our model. Images which are produced by cVAE-GAN are more diverse than cLR-GAN and BicycleGAN based on the LPIPS distance. However, it is at the expense of quality. Our CEGAN model obtains significantly better quality and diversity compared to the baselines. Through the learned

FIGURE 5. Qualitative comparison results. Comparison of randomly generated samples from different methods on the Edges–shoes dataset. Under the situation of the same input, the outputs of cVAE-GAN are diverse but not realistic(marked by red rectangle). The cLR-GAN produces realistic images while it is at the expense of diversity(marked by blue rectangle). BicycleGAN generates better outputs than the two methods, but the results from our CEGAN show the most realistic and diversest results.

TABLE 1. Quantitative evaluation metrics. Results of each method on the Edges–shoes/handbags, Map–satellite datasets and Outdoor night–day images datasets. For both metrics, the higher the better.

Methods	Edges-shoes		Edges-handbags		Map-satellite		Night-day	
	AMT Fooling %	LPIPS Distance	AMT Fooling %	LPIPS Distance	AMT Fooling %	LPIPS Distance	AMT Fooling %	LPIPS Distance
cVAE-GAN	22.56±2.85	0.171±0.021	28.69±2.07	0.227±0.058	27.13±2.58	0.155±0.037	31.34±1.96	0.308±0.075
cLR-GAN	39.27±1.97	0.121±0.014	32.15±2.58	0.142±0.028	41.66±4.34	0.094±0.009	44.61±2.38	0.198±0.036
BicycleGAN	51.62±3.26	0.159±0.025	42.18±2.14	0.195±0.035	**46.85±2.29**	0.164±0.024	47.38±3.14	0.294±0.055
CEGAN	**55.12±2.34**	**0.178±0.032**	**48.25±1.84**	**0.234±0.039**	42.17±3.27	**0.169±0.028**	**58.15±1.24**	**0.327±0.062**

latent space, we can makes full use of latent code that is distilled from the possible multiple outputs and hence obtains the multiple modes of diversity in the experiments, relatively.

E. ANALYSIS OF LATENT SPACE ADVERSARIAL LEARNING

In order to demonstrate the benefits of distinguishing real images and fake images in latent space, we have conducted

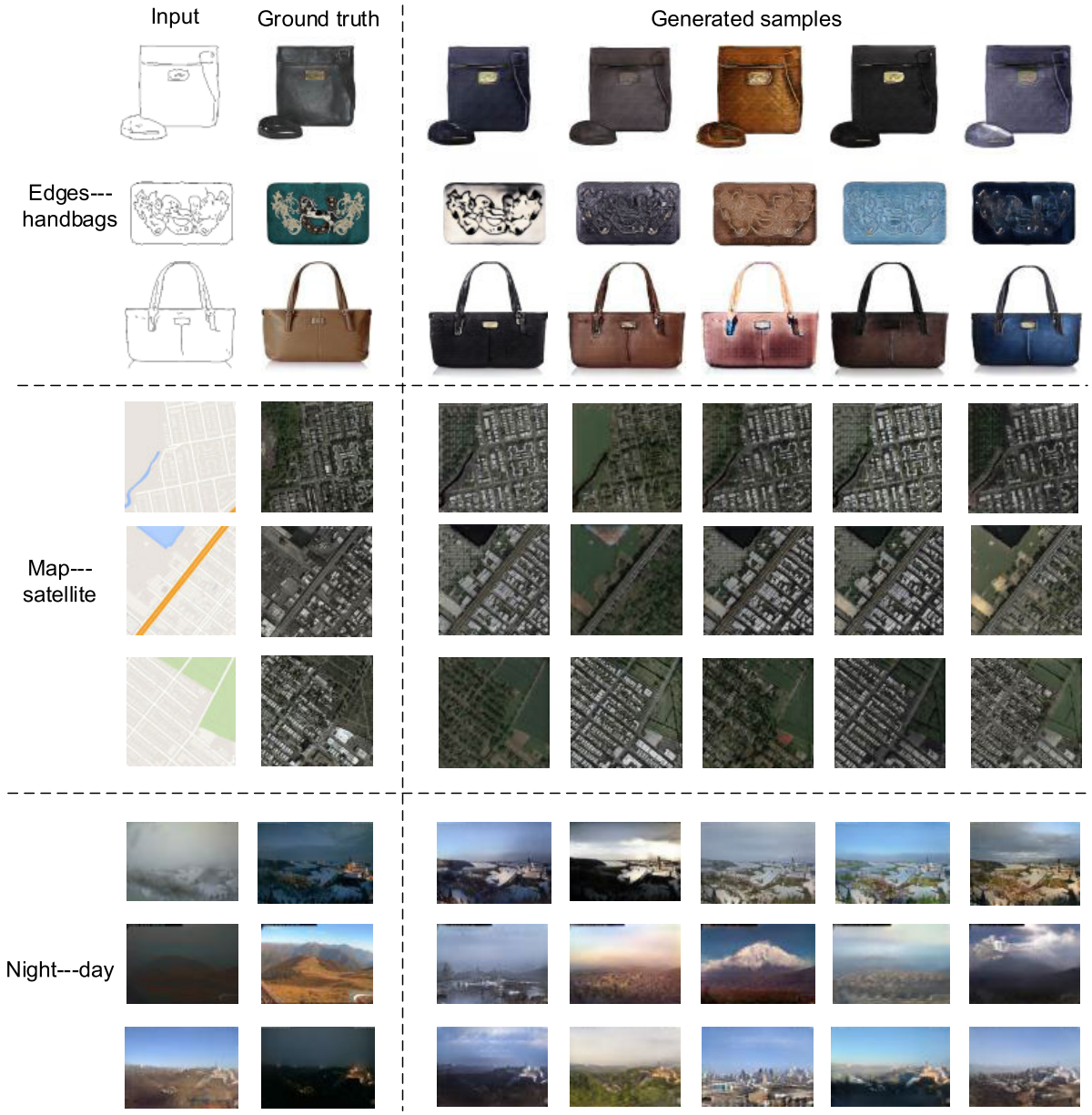

FIGURE 6. Example Results. We show example results of our model CEGAN on Edges—handbags, Map—satellite and Outdoor night—day images datasets. The left column shows the input. The second shows the ground truth column. The final five columns show randomly generated samples of our modal.

TABLE 2. Quantitative evaluation metrics. Results of each method on the Edges-shoes/handbags, Map-satellite, Outdoor night-day and Labels-facades datasets. Lower FID values mean better image quality and diversity.

Methods	Edges-shoes	Edges-handbags	Map-satellite	Night-day	Labels-facades
cVAE-GAN	0.678	0.768	1.097	2.074	1.204
cLR-GAN	0.724	0.895	1.186	1.957	1.157
BicycleGAN	0.412	**0.457**	1.073	1.429	0.924
CEGAN	**0.397**	0.552	**0.895**	**1.322**	**0.854**

more compared experiments with BicycleGAN method. Figure 7 shows the qualitative comparison results with BicycleGAN on Map-satellite, Labels-facades and Outdoor night-day datasets. Table 2 presents the quantitative FID evaluations results.

Both the quantitative and qualitative compared result verified the importance of distinguishing the real images and fake images in the latent space. As shown in Figure 7, the images generated by BicycleGAN are relatively blurred and smooth. The same phenomenon on Labels-facades dataset, we can

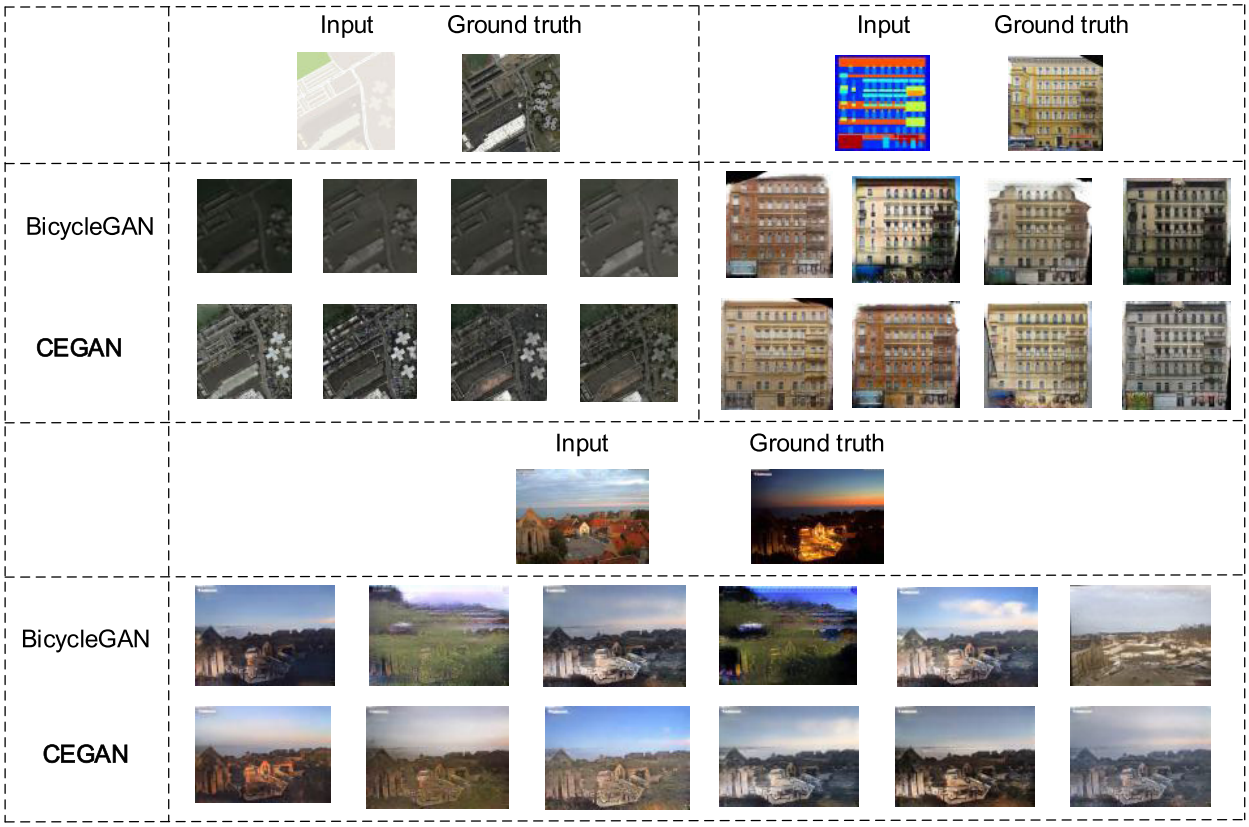

FIGURE 7. More qualitative comparison results. Comparison of randomly generated samples from BicycleGAN and our CEGAN on Map-satellite, Labels-facades and Outdoor night-day image datasets.

FIGURE 8. Ablation studies results. Samples from the model No-image-rec and No-latent-rec.

see that BicycleGAN produces images with destruction of pixel, redundancy and noise. The reason may be that it mainly considers the error relationship between the generated image and the noisy image. While through latent space learning, we can transform high-dimensional image space into low-dimensional potential space. The obtained latent code can capture the main semantic information and alleviate the impact of noise and redundancy features to some extend. The low-dimensional latent code adversarial learning is more conducive to the network training. So the generative image can be more realistic.

F. ABLATION STUDY

In this experiment, in order to understand the necessity of each individual model component to the overall performance,

we take the ablation studies by comparing the generated image quality on Edges—shoes and Map—satellite datasets. As shown in Figure 8, we mainly measure the effect of reconstruction loss to the model performance. If not mentioned otherwise, the hyper-parameters are the same in the below models.

No-Image-Rec: The loss function of this model becomes: $l_{GAN}+l_{recon}^{latent}+l_{KL}$. Note that other techniques proposed in this paper are still employed. The architecture is described as the same as Figure 3, but trained without l_1 image reconstruction loss.

No-Latent-Rec: The loss function of this model becomes: $l_{GAN}+l_{recon}^{image}+l_{KL}$. Same architecture as Figure 3, but omits the l_1 latent reconstruction loss.

Compared the generated samples, we conclude that both the image reconstruction loss and the latent reconstruction loss are important to the overall performance. The model with no-image-rec generates images which are worse quality. There are many blurs and ghosts(e.g., the first row). Again, due to l_1 loss constrains the relationship between image pixels, it makes the generated images more realistic. Without latent-rec loss, the reconstruction quality of generated samples still drops. The image texture details have been neglected. While adding the image-rec loss, the quality has improved significantly.

Table 3 shows the FID scores of each method, lower FID values mean better image quality and diversity. Both no-image-rec and no-latent-rec methods produce very worse image quality according to FID values. In general, the experiment results confirm that $l_{recon}^{latent}+l_{recon}^{image}$ offers some advantages to the final performance.

TABLE 3. Ablation studies quantitative evaluation metrics. Results of each method on the Edges—shoes and Map—satellite datasets. Lower FID values mean better image quality and diversity.

Data sets	Edges-shoes	Map-satellite
No-image-rec	0.954	1.115
No-latent-rec	0.587	0.994
CEGAN	**0.397**	**0.895**

VI. CONCLUSION

In this paper, a novel image-to-image translation model named Consistent Embedded Generative Adversarial Networks (CEGAN) is proposed to generate both realistic and diversity images. This method captures the full distribution of potential multiple modes of results by enforcing tight connections between the latent space and the real image space. Particularly, to alleviate the impact of the redundancy and noise in generated images, unlike other GANs, the discriminator in our model distinguish the real images and fake images in the latent space. Empirical experimental results showed our method is significantly better than several well-established image generation approaches.

REFERENCES

[1] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. ECCV*, 2016, pp. 649–666.
[2] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Proc. NIPS*, 2017, pp. 465–476.
[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE CVPR*, Jul. 2017, pp. 5967–5976.
[4] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "CartoonGAN: Generative adversarial networks for photo cartoonization," in *Proc. IEEE CVPR*, Jun. 2018, pp. 9465–9474.
[5] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Trans. Graph.*, vol. 35, no. 4, p. 110, 2016.
[6] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. ECCV*, 2016, pp. 577–593.
[7] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE CVPR*, Jun. 2016, pp. 2536–2544.
[8] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. IEEE CVPR*, Jul. 2017, pp. 4076–4084.
[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, X. Bing, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.
[10] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. ECCV*, 2014, pp. 184–199.
[11] M. F. Mathieu, J. J. Zhao, J. Zhao, P. Sprechmann, A. Ramesh, and Y. LeCun, "Disentangling factors of variation in deep representation using adversarial training," in *Proc. NIPS*, 2016, pp. 5047–5055.
[12] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," in *Proc. NIPS*, 2015, pp. 1486–1494.
[13] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. ICIP*, 2016, pp. 97–108.
[14] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. ICCV*, 2017, pp. 2868–2876.
[15] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. ICML*, 2017, pp. 1857–1865.
[16] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE CVPR*, Oct. 2017, pp. 2223–2232.
[17] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *Proc. IEEE CVPR*, Jul. 2017, pp. 6836–6845.
[18] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. NIPS*, 2016, pp. 469–477.
[19] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba, "Cross-modal scene networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2303–2314, Oct. 2018.
[20] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *Proc. ICIP*, 2016, pp. 1–14.
[21] J. Lin, Y. Xia, T. Qin, Z. Chen, and T.-Y. Liu, "Conditional image-to-image translation," in *Proc. IEEE CVPR*, Jun. 2018, pp. 5524–5532.
[22] X. Huang, M.-Y. Liu, S. J. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. ECCV*, 2018, pp. 179–196.
[23] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. NIPS*, 2016, pp. 2234–2242.
[24] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. ICCV*, 2017, pp. 2813–2821.
[25] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, "Wasserstein auto-encoders," in *Proc. ICLR*, 2018, pp. 1–18.
[26] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, "VEEGAN: Reducing mode collapse in gans using implicit variational learning," in *Proc. NIPS*, 2017, pp. 3308–3318.
[27] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proc. NIPS*, 2016, pp. 658–666.

[28] J. Yang, A. Kannan, D. Batra, and D. Parikh, "LR-GAN: Layered recursive generative adversarial networks for image generation," in *Proc. ICLR*, 2017, pp. 1–21.
[29] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE CVPR*, Jun. 2018, pp. 8789–8797.
[30] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked generative adversarial networks," in *Proc. CVPR*, Jul. 2017, pp. 5077–5086.
[31] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: https://arxiv.org/abs/1411.1784
[32] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE CVPR*, Jun. 2018, pp. 8798–8807.
[33] K. Sohn, X. Yan, and H. Lee, "Learning structured output representation using deep conditional generative models," in *Proc. NIPS*, 2015, pp. 3483–3491.
[34] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. CVPR*, Jul. 2017, pp. 2107–2116.
[35] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proc. ICLR*, 2017, pp. 1–14.
[36] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. CVPR*, Jul. 2017, pp. 3722–3731.
[37] S. Benaim and L. Wolf, "One-sided unsupervised domain mapping," in *Proc. NIPS*, 2017, pp. 752–762.
[38] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. NIPS*, 2017, pp. 700–708.
[39] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proc. ICCV*, Oct. 2017, pp. 1520–1529.
[40] A. Bansal, Y. Sheikh, and D. Ramanan, "PixeLNN: Example-based image synthesis," in *Proc. ICLR*, 2018, pp. 1–10.
[41] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling," in *Proc. NIPS*, 2016, pp. 82–90.
[42] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Proc. ICML*, 2017, pp. 1–4.
[43] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.
[44] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICIP*, 2015, pp. 1–15.
[45] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *Proc. IEEE CVPR*, Jun. 2014, pp. 192–199.
[46] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Proc. ECCV*, 2016, pp. 597–613.
[47] S. Xie and Z. Tu, "Holistically-nested edge detection," *Int. J. Comput. Vis.*, vol. 125, nos. 1–3, pp. 3–18, Dec. 2017.
[48] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays, "Transient attributes for high-level understanding and editing of outdoor scenes," *ACM Trans. Graph.*, vol. 33, no. 4, p. 149, 2014.
[49] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. CVPR*, Jun. 2016, pp. 3213–3223.
[50] R. Tyleček and R. Šára, "Spatial pattern templates for recognition of objects with regular structure," in *Proc. GCPR*. Springer, 2013, pp. 364–374.
[51] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE CVPR*, Jun. 2018, pp. 586–595.
[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[53] D. Dowson and B. Landau, "The Fréchet distance between multivariate normal distributions," *J. Multivariate Anal.*, vol. 12, no. 3, pp. 450–455, 1982.

FENG XIONG received the B.Eng. degree from the Hubei University of Technology, Wuhan, China, in 2017. He is currently pursuing the M.S. degree with Xidian University, Xi'an, China. His current research interests include pattern recognition and machine learning.

QIANQIAN WANG received the B.Eng. degree in communication engineering from the Lanzhou University of Technology, China, in 2014. She is currently pursuing the Ph.D. degree in communication and information system with Xidian University, China. Her research interests include pattern recognition, dimensionality reduction, sparse representation, and face recognition.

QUANXUE GAO received the B.Eng. degree from Xi'an Highway University, Xi'an, China, in 1998, the M.S. degree from the Gansu University of Technology, Lanzhou, China, in 2001, and the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2005. He was an Associate Researcher with the Biometrics Center, The Hong Kong Polytechnic University, Hong Kong, from 2006 to 2007. He is currently a Professor with the School of Telecommunications Engineering, Xidian University, and also a Key Member of the State Key Laboratory of Integrated Services Networks. His current research interests include pattern recognition and machine learning.