Image translation with dual-directional generative adversarial networks

Congcong Ruan | Liuchun Yuan | Haifeng Hu | Dihu Chen

School of Electronics and Information Technology, Sun Yat-Sen University, Guangzhou, People’s Republic of China

Correspondence
Dihu Chen, School of Electronics and Information Technology, Sun Yat-Sen University, Guangzhou, People’s Republic of China. Email: stscdh@mail.sysu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 61673402, 61733270, 60802669; Natural Science Foundation of Guangdong Province, Grant/Award Numbers: 2017A030310129, 2016B010123005, 2017B090900005, 2019B010140002; Science and Technology Program of Guangzhou of China, Grant/Award Numbers: 201704020180, 201604020024; Fundamental Research Funds for the Central Universities of China

Abstract
Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between input images and output images. However, due to the unstable training and limited training samples, many existing GAN-based works have difficulty in producing photo-realistic images. Herein, dual-directional generative adversarial networks are proposed, which consist of four adversarial networks, to produce images of high perceptual quality. In this framework, self-reconstruction strategy is used to construct auxiliary sub-networks, which impose more effective constraints on encoder-generator pairs. Using this idea, this model can increase the use ratio of paired data conditioned on the same dataset and obtain well-trained encoder-generator pairs with the help of the proposed cross-network skip connections. Moreover, the proposed framework not only produces realistic images but also addresses the problem where condition GAN produces sharp images containing many small, hallucinated objects. Training on multiple supervised datasets, convincing evidences are shown to prove that this model can achieve compelling results by latent learning a common feature representation. Qualitative and quantitative comparisons against other methods, demonstrate the effectiveness and superiority of the method.

1 | INTRODUCTION

Many problems in image processing, computer graphics, and computer vision can be posed as ‘translating’ an input image into a corresponding output image. Translating an image from source domain $X$ to target domain $Y$ is equal to learning a data distribution transformation between domains. We define the problem of image transformation from a scene into another as image-to-image translation.

In the field of image translation, prior works use convolutional neural nets (CNNs) to achieve image-to-image translation via variational autoencoders (VAEs) [1,2]. They use CNNs to produce images, which are relatively blurry, by minimizing Euclidean distance between generated images and real images [3,4]. This is because Euclidean distance is minimized by averaging all plausible outputs, which causes blurring. Nevertheless, cascaded refinement networks [5] propose a new method to prove that CNNs can also synthesize more realistic images than generative adversarial networks (GANs). Generally speaking, in image processing and graphics tasks, GANs proposed by GoodFellow [6] are becoming a popular framework on problems where the output is highly detailed or photographic. With the help of GANs, recent research in computer vision, image processing, and graphics has produced powerful translation systems in the supervised setting, where example image pairs $\{x, y\}$ are available, for example, the studies by Eigen and Fergus, Isola et al., Johnson et al., Long et al., and Ruan et al. [7–11]. Nevertheless, these works still have difficulty in producing photo-realistic images like real-world images. Compared with previous works, the study by Isola et al. [8] makes great progress but generated images still have a large gap with the ground truth, demonstrating that parameters of the model are not been well learned. For an ideal encoder-generator pair for image translation, its encoder is supposed to have the ability to extract distilled content information as style information of input images is of no use to produce output images and even affects the performance of generators. The ideal generator should be a nice domain-specific render that can produce photo-realistic images from the learned feature. Actually, GANs are hard to have a stable training and lack of necessary constraints so that the
encoder extracts feature representations that contain useless style information and may lose the key content information, resulting in bad results including many small, hallucinated objects. On the other hand, even if the ideal generator is available, content information mixed with the redundant style information makes it difficult for generators to have a good rendering effect on generated images, resulting in blurry images or defective images.

To address the issue above, we seek a translation network that can learn the parameters well and produce photo-realistic images. There is no doubt that enlarging datasets is an efficient method. Nevertheless, large paired datasets are difficult to obtain and the resource of paired datasets is extremely limited. Therefore, we think that increasing the use ratio of data is an effective strategy that can produce impressive results conditioned on the same dataset. For the task of supervised image-to-image translation, we assume that there is some underlying relationship between domains. For example, as shown in Figure 1, domain X has a close relation with corresponding domain Y. Specifically, paired images differ in surface appearance while both are renderings of the same underlying structure and share the content information of images. To make use of this character, we construct a crossing framework, including two main sub-networks $X \rightarrow Y, Y \rightarrow X$ and two auxiliary sub-networks $X \rightarrow X, Y \rightarrow Y$, to facilitate mapping learning between domains. In our framework, it implements the task of image translation and image reconstruction for both domains at the same time. We feed feature representations of both domains to domain-specific generators so that our encoders can extract common feature representations of input images, that is, content information. Meanwhile, our generators tend to ignore the redundant style information and act as domain-specific renderers. Taking advantage of the special cross-domain framework, our model can obtain better encoder-generator pairs using two main sub-networks: $X \rightarrow Y, Y \rightarrow X$ together with two auxiliary sub-networks: $X \rightarrow X, Y \rightarrow Y$. Thus the proposed network is named dual-directional generative adversarial network (GANs). Besides, inspired by U-Net [12], cross-network skip connections are proposed in our framework to enhance the performance of generators since images of both domains have a similar underlying relationship. Unlike the study by Liu and Tuzel [13], our core idea and principle are dedicated to solving the issue of parameters learning in supervised learning, which is quite different from them. Besides, the key point is that our model does not use the weight-sharing strategy between encoders of different domains, and it is the same to GANs. For more details about our model, please see Section 3.

Herein, we explore GANs to achieve better results in the field of image-to-image translation using supervised learning. The main contribution of this work is that we propose a novel framework composed of four adversarial networks, which is called dual-directional GANs. The proposed framework utilizes cross-network skip connections and self-reconstruction to facilitate mapping learning between domains. Benefitting from the special framework, our encoders have the ability to extract intact content information and our generators have a nice rendering effect on generated images. Second, the proposed model can not only produce more realistic images but also figure out the problem where condition GAN produces sharp images containing many hallucinated objects. Third, extensive experiments are conducted to show that the proposed model is able to yield better results than some existing works. Qualitative and quantitative comparisons against other methods, demonstrate the effectiveness and superiority of our idea.

2 | RELATED WORKS

In our work, we succeed in achieving image-to-image translation with multiple technologies, including GANs, image translation, multi-network learning and so on. Next, we will introduce related works about three key points in detail.

2.1 | Generative adversarial networks (GANs)

Generating impressive images is still a challenging problem until the rapid development of the GAN. GANs have been
proposed by Goodfellow et al. It simultaneously trains two models: 1) the generator G tries to learn the mapping from source domain $X$ to target domain $Y$, and 2) the discriminator D tries to distinguish the generated images from real images. The key to success of GANs is the novel framework that utilizes an adversarial loss to force generated images to be indistinguishable from ground truth. With the rapid advancement of the GANs, many works have succeeded in achieving impressive results in image generation. For example, DCGAN [14] utilizes convolution-deconvolution network to convert random noise to meaningful images. CGAN [15] succeeds in producing more realistic images by feeding conditional variable (e.g., the class label) to the input of network. In our work, we also use GANs to achieve image-to-image translation because of its superiority in image generation.

### 2.2 Image-to-image translation

Converting one image to another can be considered the task of image-to-image translation. The idea of image-to-image translation goes back at least to Hertzmann et al.’s Image Analogies [16], which uses a nonparametric texture model [17] from a single input–output training image pair. Prior works, like Gatys et al. [18], apply CNNs to produce new images of high perceptual quality that combine the image content and style of natural images. More recent approaches use GANs to train their models and achieve impressive results. Isola et al. [8] figure out a general network called ‘pix2pix’ to achieve image translation using paired training datasets. They use an image-conditional GAN to learn a mapping from source images to target images. The image-conditional models have tackled image prediction from a normal map [19], future frame prediction [20], product photo-generation [21,22], and image generation from sparse annotations [23,24]. It is interesting that Chen et al. [5] utilize CNNs to make great progress in image translation. Unlike contemporary works, their approach can produce high-definition images by use of cascaded refinement networks without adversarial training. In addition to supervised learning, a large amount of works have achieved fruitful results in unsupervised setting. For example, DiscoGAN [25] proposes a method based on GANs that learns to discover relations between different domains. Yi et al. [26] develop a novel dual-GAN mechanism to learn both mappings between domains. The primal GAN learns to translate images from domain $X$ to domain $Y$, while the dual GAN learns to convert the task. The closed loop made by the primal and dual tasks allows images from either domain to be translated and then reconstructed. The key idea is similar to CycleGAN [27] as a concurrent work and both of them use the closed loop to acquire good results in unsupervised setting. Similar to our work, both of them can learn to translate images between domains in dual directions. Different from the cycle-reconstruction strategy used in CycleGAN and DualGAN [26], our self-reconstruction is reconstructed from feature space instead of image space. Besides, our framework does not use the closed loop, and it is quite different from them. Specifically, DualGAN utilizes the cycle-reconstruction strategy to translate images $X \rightarrow Y$ and then reconstructs input images with generated images as the input to achieve $Y \rightarrow X$, which is a close-loop network and based on RGB images. Contrary to it, our model is an open-loop network, which consists of four adversarial networks to achieve image translation and image reconstruction using common feature representations instead of RGB images used in DualGAN. It is notable that we do not reuse the generated images to reconstruct input images. That is to say that we do not apply cycle-reconstruction strategy used by DualGAN in our model.

### 2.3 Multi-network learning

In the field of image generation, several works have already implemented multi-network learning to tackle multi-domain problems. Zhu et al. [27] use the idea of cycle consistency to construct two GANs that can learn both mappings between domains using unpaired training datasets. CoupledGANs [13] use two parallel networks to learn a common representation across domain with weight-sharing strategy. Like CoupledGANs, Liu et al. [28] use two partly weight-sharing networks, which combine VAEs and GANs, to learn unsupervised image-to-image translation. Note that weight-sharing strategy plays an important role in cross-domain unsupervised learning. Different from them, our model does not use the weight-sharing strategy between encoders of different domains and it is the same to GANs. It is worth saying that multi-network learning is becoming a popular method in recent works. For example, Lin et al. [29] decompose images into two parts: the content and style of images, respectively. To translate the image to target domain, they recombine its content code with the style code of another image from target domain by use of image-reconstruction strategy. MUNIT [30] proposes a similar idea to achieve multi-modal images translation, which implements the content code reconstruction as well as the style code. Like cd-GAN [29], MUNIT allows users to control the style of translation outputs by providing an example style image. Karras et al. [31] proposes an unsupervised image-to-image translation framework which maximizes a domain-specific variational information bound and learns the target domain-invariant representation of the two domain. Lee et al. [32] present an approach based on disentangled representation for producing diverse outputs without paired training images. Gonzalez et al. [33] proposes a framework based on BicycleGAN [34] to achieve multimodal image translation by separating the internal representation into two parts: the shared content part and the exclusive style part. The content part is later concatenated with the style part to achieve multi-modal transformations. Different from it, our model utilizes the proposed framework to force encoders to extract the distilled content information and discard the useless style information. Our generators are domain-specific renderers that only take content representations as input to synthesize images belonging to corresponding domains. Moreover, we do not impose more constraints on feature space with L1 norm loss,
FIGURE 2 The framework of our dual-directional adversarial network. Purple blocks represent input and output domains. Green blocks represent encoder $E$, generator $G$ and discriminator $D$. Yellow blocks represent the feature representations obtained by encoders. Black arrow lines represent the direction of information flow and red arrow lines represent skip connections between encoder $E$ and generator $G$. It is assumed that paired images have the same underlying structure, so we take no actions to link feature representations $Z_X$ and $Z_Y$ in our network. Note that the input is real training images, and the output is synthesized images generated by our model.

| Mapping | $M_Y: X \rightarrow Y$ | $M_X: Y \rightarrow X$ | $F_Y: Y \rightarrow Y$ | $F_X: X \rightarrow X$ |
|---------|-------------------------|-------------------------|-------------------------|-------------------------|
| Sub-network | $\{E_X, G_Y, D_Y\}$ | $\{E_Y, G_X, D_X\}$ | $\{E_Y, G_Y, D_Y\}$ | $\{E_X, G_X, D_X\}$ |

which is used in the study by Gonzalez et al. [33]. Among these works, most use the main idea of cycle consistency to achieve image translation in essence.

3 | METHOD

In this section, we give a general overview to our model and share the formulation of our novel framework. The loss functions of our model are explained in the end.

3.1 | Our network

As shown in Figure 2, our framework is based on autoencoders and GANs. It consists of two main sub-networks $M_Y: X \rightarrow Y$, $M_X: Y \rightarrow X$ and two auxiliary sub-networks $F_Y: Y \rightarrow Y$, $F_X: X \rightarrow X$. Every sub-network consists of three parts: including one domain encoder $E$, one domain generator $G$ and one domain discriminator $D$. In our framework, there are six blocks in total: encoder $E_X$ for domain $X$ and $E_Y$ for domain $Y$, generator $G_X$ for domain $X$ and $G_Y$ for domain $Y$, discriminator $D_X$ for domain $X$ and $D_Y$ for domain $Y$. The encoder-generator pair $\{E, G\}$ constitutes a sub-network that first encodes source images to shared code via $E$ and then decodes the shared code to produce target images via $G$. $D_X$ and $D_Y$ aim to distinguish generated images from real images of corresponding domain. It is worth noting that main sub-network $M_Y$ has different encoders from auxiliary sub-network $F_Y$. Also, it is same to $M_X$ and $F_X$. For example, main mapping $M_Y: X \rightarrow Y$ has the same generator $G_Y$ and discriminator $D_Y$ with auxiliary mapping $F_Y: Y \rightarrow Y$. However, they have different encoders, specifically $E_X$ for mapping $M_Y, E_Y$ for mapping $F_Y$. To make this easy to understand, we list every part of sub-networks in Table 1. More specifically, both of our encoders are constructed with identical network architecture, which is composed of eight convolution blocks to extract common feature representations. Each block has the form of Convolution-InstanceNorm-LReLU except the first block without InstanceNorm. We use $4 \times 4$ convolutions with stride two to downsample the input image to 512 feature maps of size $1 \times 1$. Corresponding generators utilize eight fractionally strided convolutions blocks with stride half, which use modules of the form Deconvolution-InstanceNorm-ReLUs except the last block using Deconvolution-tanh, to make the input and output have exactly the same size. Our discriminator is an image-conditional convolutional discriminator that takes both of input and output images as the input of networks. Note that our discriminators do not adopt PatchGAN [35]. Instead, we take the entire images as the input of discriminators. The
discriminator first utilizes four convolutions blocks with stride two to downsample the input image to 512 dimensions and then uses one convolution block with stride one to output the score. But why does our model work? Next we will explain it in detail.

The traditional method like ‘pix2pix’, is a single network and the model is lack of necessary constraints on their encoders and generators. In our work, our goal is to learn mappings between two domains $X$ and $Y$ given paired training samples $\{x_i\} \in X$ and $\{y_i\} \in Y$. As mentioned in Section 1, a single encoder-generator pair $\{E_X, G_Y\}$ cannot ensure that $E_X, G_Y$ act as an ideal encoder and generator, respectively. Considering that paired samples $x_i$ and $y_i$ have the same content information, we take both feature representations $E_X(x_i)$ and $E_Y(y_i)$ as the input of corresponding generators $G_X$ or $G_Y$ to train our model. For example, as shown in Figure 3, $G_Y$ takes the feature representations $E_X(x_i), E_Y(y_i)$ to produce transformed images $G_Y(E_X(x_i)), G_Y(E_Y(y_i))$ under supervision using a L1 norm loss with real images $y_i$. To minimize L1 loss, feature representation $E_X(x_i)$ will be gradually close to $E_Y(y_i)$. The final result is that our encoders are forced to extract common feature representations, that is, content information. In supervised setting, the content information of domain $X$ is similar to that of domain $Y$. Meanwhile, this training strategy makes $G_Y$ insensitive to style information in $E_X(x_i)$ or $E_Y(y_i)$ by updating parameters step by step. It is because that domain $X$ has different style information from domain $Y$. If our generators do not ignore the style information in $E_X(x_i)$ or $E_Y(y_i)$, the L1 loss between $G_Y(E_X(x_i)), G_Y(E_Y(y_i))$ and $y_i$ will be enlarged and it violates the principle of minimizing loss. Hence, our generators have a nice rendering effect on generated images. In a word, the goal of constructing auxiliary sub-networks is to obtain well-trained encoder-generator pairs. To have a better model, we implement a similar operation on $G_X$. On the basis of $M_Y$ and $F_Y$, we introduce $M_X$ and $F_X$ to facilitate mapping learning and make our encoders $E_X, E_Y$ act as an ideal encoder, which has the ability to extract intact content information without style components. As a consequence of constructing four sub-networks, the use ratio of data is increased during the training time. Benefiting from the special framework, our model can obtain better encoders, generators and produce more realistic images.

In addition to auxiliary sub-networks, we also add skip connections to generators. Every sub-network consists of autoencoders and GANs. Input images are passed through a series of layers until they are progressively downsampled to shared feature, and then GANs decode it to output images that have the same size $256 \times 256$ as input images. Considering the small dimension of shared feature, there is a great deal of low-level information discarded when the input is passed through

---

**Figure 3** As the training goes on, the feature representation extracted by $E_X$ is gradually approaching that extracted by $E_Y$. Finally, $E_X$ and $E_Y$ get the similar feature, that is, content information.
The objectiv e of a conditional GAN can be expressed as

\[ \text{Objectives} \]

\[ l_{\text{GAN}}(E, G, D) = \log D(x, y) + \log[1 - D(x, G(E(x)))] \]

Where encoder \( E \) converts the input \( x \) to latent codes \( E(x) \). Discriminator \( D \) tries to minimize this objective by forcing \( D(x, y) \) to approach one and \( D(x, G(E(x))) \) to approach zero, that is \( D(x, y) \approx 1, D(x, G(E(x))) \approx 0 \). It tries to distinguish real images \( y \) from synthesized images \( G(E(x)) \) by playing a max-min game with encoder \( E \) and generator \( G \). Generator \( G \) tries to produce images \( G(E(x)) \) that look similar to images from domain \( Y \), that is \( D(x, G(E(x))) \approx 1 \). In a word, \( G \) tries to minimize this objective against an adversarial \( D \) that tries to maximize it, that is, \( G^* = \arg \min_{G} \max_{D} l_{\text{GAN}}(E, G, D) \). In our model, it includes two main mappings \( M_X: X \rightarrow Y, M_Y: Y \rightarrow X \) and two auxiliary mappings \( F_X: X \rightarrow Y, F_Y: Y \rightarrow X \). All of them have a similar adversarial loss as conditional GAN.

Another term is L1 norm loss, which prior works have found it beneficial to combine the GAN objective with. L1 loss forces CNNs to learn mappings between domains so that generated images are closer to the ground truth. L1 objective can be expressed as

\[ l_{L1}(E, G) = \|y - G(E(x))\|_1 \]

In conclusion, our final objective functions consist of two terms: L1 loss and adversarial loss. L1 loss facilitates our model to learn the appropriate mapping between domains and model the data distribution of datasets. Adversarial loss is introduced to regularize the parameters to make generated images more realistic like real-world images. In our model, we utilize conditional GAN [8], which takes generated images and input images together as the input of discriminators to produce images of high perceptual quality. The final cost functions of our encoder \( E_X, E_Y \) and generator \( G_X, G_Y \) can be expressed as

\[ M_X : l_{E_X, G_X} = -\log D_X(y, G_X(E_Y(y))) + \lambda \|y - G_X(E_Y(y))\|_1 \]

\[ M_Y : l_{E_Y, G_Y} = -\log D_Y(x, G_Y(E_X(x))) + \lambda \|y - G_Y(E_X(x))\|_1 \]

\[ F_X : l_{E_X, G_X} = -\log D_X(y, G_X(E_Y(x))) + \lambda \|y - G_X(E_Y(x))\|_1 \]

\[ F_Y : l_{E_Y, G_Y} = -\log D_Y(x, G_Y(E_X(y))) + \lambda \|y - G_Y(E_X(y))\|_1 \]

The hyper-parameter \( \lambda \) is the weight used to balance L1 loss with adversarial loss. Our encoders \( E \) and generators \( G \) try to minimize loss functions \( l_{E_X, G_X}, l_{E_Y, G_Y} \) by forcing generated images close to the ground truth. For example, in case of sub-network \( F_Y \), \( D_Y(x, G_Y(E_Y(y))) \) approaches 1 and \( \|y - G_Y(E_Y(y))\|_1 \) approaches 0. The cost functions of our discriminator \( D_X, D_Y \) can be expressed as

\[ M_X : l_{D_X} = \log D_X(y, x) + \log[1 - D_X(y, G_X(E_Y(y)))] \]

\[ M_Y : l_{D_Y} = \log D_Y(x, y) + \log[1 - D_Y(x, G_Y(E_X(x)))] \]

\[ F_X : l_{D_X} = \log D_X(y, x) + \log[1 - D_X(y, G_X(E_Y(x)))] \]

\[ F_Y : l_{D_Y} = \log D_Y(x, y) + \log[1 - D_Y(x, G_Y(E_X(y)))] \]

To minimize cost functions of \( D_X, D_Y \), discriminators \( D_X, D_Y \) are forced to distinguish generated images from real images through parameters learning.

4 | EXPERIMENTS

In this section, we divide our experiments into three parts to introduce. We start with introducing various datasets and baselines used in our experiments. Then, qualitative comparisons on perceptual effect are carried out to show that dual-directional GANs has the ability to generate more realistic images than other methods. In the end, we use the same metrics as 'pix2pix' [8] to evaluate the performance of the proposed method. The quantitative comparisons against other methods are conducted to show the superiority of our model. Besides, ablation study is conducted to further prove
the effectiveness of dual-directional GANs. To optimize our network, we apply the Adam optimizer [36] where the learning rate is set to 0.0002 and momentums are set to 0.5 and 0.999. In our experiments, we set self-reconstruction weight $\lambda = 100$ and train our model for 200 epochs with the minibatch size of 1. It is notable that using cross-network skip connections should obey one principle. As shown in Figure 2, generator $G_Y$ shares the information from encoder $E_X$ and generator $G_X$ shares the information from encoder $E_Y$.

4.1 Datasets and baselines

To prove the validity of dual-directional GANs, we test the method on a variety of datasets, including Cityscapes dataset [37], GTA5 dataset [38], sketch-photo [39] and Google Maps [8]. Note that all images used in our experiments, including paired training images, val images, generated images and so on, have the same size 256 $\times$ 256 except CRN [5], which produces images of 512 $\times$ 256.

Cityscapes dataset is a challenging real-world dataset. It contains urban street images collected from a moving vehicle captured in 50 cities around Germany and neighbouring countries. Cityscapes dataset has 2975 paired training images and 500 val images. We conduct both qualitative and quantitative comparisons on Cityscapes dataset to prove the effectiveness of dual-directional GANs.

GTA5 dataset consists of 24,966 densely labelled RGB images/video frames, containing 19 classes that are compatible with the Cityscapes dataset. In our experiments, we select the first 5000 paired images as experimental datasets for performance comparison. To avoid the similarity and continuity of training images, 2500 images of even numbers are chosen as the training set and the remaining images serve as the val set.

Sketch-photo dataset contains 995 paired training images and 199 val images. It will be mainly used in the experiment of qualitative comparison.

Google Maps dataset consists of 1096 paired training images and 1098 val images that are scraped from Google Maps. It is used on the map $\rightarrow$ aerial photo setting.

In our experiments, we carry out several performance comparisons against some existing works on multiple datasets. The baselines include some compelling works, that is, 'pix2pix' [8], DualGAN [26], CycleGAN [27], MUNIT [30], CRN [5] and PAN [40]. It is notable that MUNIT is a state-of-the-art approach on synthesizing diverse and realistic images. Like edGAN [29], MUNIT has the ability to control the style of translation outputs by providing an example style image. Thus it is suitable to the task of image translation. For fair comparisons, we replace the random style code with the style code of given style images to achieve the task of image translation. Besides, according to instructions in their repository, we use the pre-trained 512 $\times$ 256 CRN model to produce images of 512 $\times$ 256 for comparisons. DualGAN is referred to as Dual in our experiments for convenience.

4.2 Qualitative comparison

In the experiment, we conduct several qualitative comparisons between our model and other methods. For fair comparisons, we implement the task of image-to-image translation on the same training dataset but with different networks. Figure 4 shows synthesized images produced by different models on Cityscapes dataset. As we can see, images generated by our model are much better than images generated by 'pix2pix' model in perceptual quality. At the local part of images generated by 'pix2pix' model, there is much noise so that they look blurry. Compared with them, images generated by our model look sharper and more realistic on perceptual intuition. Moreover, compared with the state-of-the-art approach like CRN, MUNIT, our model also outperforms them. In our view, CRN is a model based on CNNs, which is the major reason for lower definition. It is worth emphasizing that CRN heavily relies on cascaded refinements of the pre-trained model at the resolution of 256 and 512. At the resolution of 1024 $\times$ 2048, CRN outperforms other models without cascaded refinements as it is difficult for models without cascaded refinements to learn so large quantity of parameters at a time. In our experiments, although images generated by CRN have a resolution of 512 $\times$ 256, which is larger than 256 $\times$ 256 used in other methods, they look inferior to the images of our model in perceptual quality, as shown in Figure 4. Furthermore, it is notable that MUNIT plays not well on complex paired datasets, for example, driving datasets like Cityscapes and GTA5 datasets. It is because that urban scenes have much complex information so that corresponding images cannot be simply decomposed into the style code and content code. MUNIT is only suitable to some simple datasets like edges $\rightarrow$ shoes [8]. Unlike those works, our model has a broad application and achieves a better effect on the task of image translation. It implements noise elimination and improves the quality of generated images. Figure 5 presents the example results on GTA5 dataset. It is observed that the images generated by our model have higher resolution and are more realistic like real-world images. More results on labels $\rightarrow$ photos can be seen in Figures 6 and 7.

As shown in Figure 8, images generated by 'pix2pix' model have another drawback that they include many small,
hallucinated objects on the task of \textit{photo} \rightarrow \textit{labels}. Isola et al. encounter the same problem therein [8], and they use quantitative classification accuracies to demonstrate that simply using L1 regression gets better scores than using a condition GAN (with/without L1 loss) on the task of \textit{photo} \rightarrow \textit{labels}. They argue that for vision problems, the goal (i.e., predicting output close to ground truth) may be less ambiguous than graphic tasks, and reconstruction losses like L1 norm are mostly sufficient. Although condition GAN can produce sharper images than those using L1 loss, it brings in hallucinated objects simultaneously. For the translation task on Cityscapes dataset and Google maps, we conduct experiments on DualGAN [26] and find it is easy for DualGAN to cause mode collapse that outputs the same images regardless of input images. Furthermore, we carry out several experiments on driving datasets and find it is difficult for most models to learn well on the task of \textit{photo} \rightarrow \textit{labels}. In our work, we still use a condition GAN with L1 loss as Equations 3 and 4, but different frameworks to produce realistic images without hallucinated objects in Figure 8. We use the proposed framework to figure out that condition GAN produces images containing many small, hallucinated objects. To our knowledge, it is the first time for conditional GAN to solve this issue on Cityscapes dataset. As we explain in Section 3, our model can utilize the novel framework to obtain much better encoder-generator pairs. Specifically, our encoder has the ability to extract intact content information and our generator has a nice rendering effect on generated images. Traditional encoder-generator pair is lack of necessary constraints so that the encoder may lose the content information of input images and the generator has a not good rendering effect on the output, resulting in blurry images including many small, hallucinated objects. More results on \textit{photo} \rightarrow \textit{labels} can be seen in Figure 9.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5.png}
\caption{Qualitative comparisons on the task of GTAS \textit{labels} \rightarrow \textit{photo}. Note that our method can produce sharper images where objects, for example, roads, cars and motorbikes, have clear profiles and colour pixels.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6.png}
\caption{Example results for the task of \textit{labels} \rightarrow \textit{photos} on sketch-photo dataset.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure7.png}
\caption{Example results on Google Maps. Compared with other methods, our model can produce images of high definition. Please zoom in as possible as you can for more details about clarity.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure8.png}
\caption{Qualitative comparisons on the task of Cityscapes \textit{photo} \rightarrow \textit{labels}. Note that 'pix2pix' model produces images that include many small, hallucinated objects. In contrast to other methods, our model can produce more realistic and precise images, which are much close to the ground truth. Please zoom in for more information about image definition.}
\end{figure}
4.3 | Quantitative evaluation

To illustrate the performance of images translation tasks, we use quantitative measures to evaluate the performance over the test sets, such as ‘FCN score’ [8], Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [41] and Visual Information Fidelity (VIF) [42]. Besides, ablation study is conducted to further prove the effectiveness of dual-directional GANs.

4.3.1 | Evaluation metrics—FCN-score

Although perceptual comparison may be the gold standard for evaluating graphical realism, we try to find an automatic quantitative measure that can provide specific evaluation data. To this end, we adopt the ‘FCN score’ from the study by Isola et al. [8], and use it to evaluate the task of semantic labels → photos on Cityscapes dataset with the standard metrics. The Cityscapes benchmark includes per-pixel accuracy, per-class accuracy, and mean class Intersection-Over-Union (Class IOU) [37]. The FCN predicts a label map for a synthesized photo, and then compared with the ground truth label, the predicted label map is given a FCN-score with standard semantic segmentation metrics. If synthesized photos are realistic enough, corresponding predicted labels will highly approach to ground truth labels and get high scores simultaneously.

4.3.2 | Performance comparison

Table 2 lists the comparison results on the task of Cityscapes labels → photos. Note that we quote partial data from the studies by Isola and coworkers [8,27] for comparison and all the baselines have the same implementation details as our method. Note that CRN heavily depends on cascaded refinements and we use the pre-trained 512 × 256 model provided by the study by Chen and Koltun [5] as the baseline in this experiment. According to the evaluation data, as shown in Table 2, our model gets the highest scores and outperforms baselines by a large margin. Compared with ‘pix2pix’ model, our model yields the highest per-pixel accuracy up to 0.73, which outperforms ‘pix2pix’ by 7 percent points. Moreover, we use more metrics (i.e., PSNR, SSIM, VIF) to evaluate performance of different models. Tables 3 and 4 show the evaluation data on Cityscapes labels → photos and labels → photos, respectively. The experimental results suggest that our model can play a better role on the task of image-to-image translation.

![Figure 9](image-url) Figure 9: Example results for the task of photo → labels on sketch-photo dataset. Note that DualGAN produces images that include small, hallucinated objects on the local part of faces.

| Table 2 | FCN-scores for different methods, evaluated on Cityscapes labels → photos |
| --- | --- | --- |
| Method | Per-pixel acc. | Per-class acc. | IOU |
| BiGAN | 0.19 | 0.06 | 0.02 |
| GoGAN | 0.40 | 0.10 | 0.06 |
| CRN | 0.38 | 0.14 | 0.08 |
| MUNIT | 0.45 | 0.11 | 0.07 |
| CycleGAN | 0.52 | 0.17 | 0.11 |
| Pix2pix | 0.66 | 0.23 | 0.17 |
| Ours | 0.73 | 0.25 | 0.19 |

| Table 3 | Performance comparison on the task of Cityscapes labels → photos |
| --- | --- | --- |
| Method | PSNR(dB) | SSIM | VIF |
| MUNIT | 11.81 | 0.258 | 0.0147 |
| CycleGAN | 14.22 | 0.326 | 0.0189 |
| CRN | 15.46 | 0.397 | 0.0199 |
| Pix2pix | 15.35 | 0.4223 | 0.0529 |
| PAN [40] | 16.06 | 0.4820 | 0.0658 |
| Ours | 17.76 | 0.4609 | 0.0664 |

| Table 4 | Performance comparison on the task of Cityscapes photos → labels |
| --- | --- | --- |
| Method | PSNR(dB) | SSIM | VIF |
| MUNIT | 12.92 | 0.546 | 0.0825 |
| CycleGAN | 16.86 | 0.613 | 0.1072 |
| Pix2pix | 18.12 | 0.6911 | 0.1123 |
| ID-cGAN [43] | 19.46 | 0.7270 | 0.1180 |
| PAN [40] | 20.67 | 0.7725 | 0.1638 |
| Ours | 20.72 | 0.7642 | 0.1692 |
4.3.3 Ablation study

In this section, we carry out several experiments to show how each component in our loss function affects the final performance. Without changing hyper-parameters, we only add or remove some component of the loss function for fair comparisons. We consider the following cases: (a) Ours(full): the full implementation of our approach including four parts: $M_X, M_Y, F_X, F_Y$. (b) Ours w/o $M_X$: On the basis of Ours(full), sub-network $M_X$ is removed. (c) Ours w/o $F_X$: On the basis of (b), sub-network $F_X$ is removed. (d) Our base: Here, we remove sub-network $F_Y$ and reserve sub-network $M_Y$. The experiments results on the Cityscapes val set are shown in Table 5. According to the evaluation data, it can be observed that each sub-network is important to obtain the full improvement in performance. With the help of sub-networks $F_Y, M_Y$, our model can obtain a better generator $G_Y$ and produce more realistic images, which get a higher score up to 0.707 in per-pixel accuracy compared with a single network $M_Y$. To further improve performance, we add sub-networks $M_X, F_X$ to train our encoders $E_X, E_Y$ together with generators $G_Y$ and $G_X$ so that our encoders have the ability to extract intact content information without style components and our generators act as ideal ones to produce images of high fidelity and definition. The experiment showcases the effectiveness of the proposed framework.

In addition, the ablation study on the cross-network skip connections is conducted to prove its effectiveness. We consider the following cases: (a) Ours(full): the full implementation of cross-network skip connections on four sub-networks: $M_X, M_Y, F_X, F_Y$. (b) Ours w/o $F_X, F_Y$: The cross-network skip connections on $F_X, F_Y$ are removed. That means the two red arrow lines in the middle of Figure 2 are removed from our framework. In our generators, we replace the removed feature with the special maps, which consist of zero and have the same size as the removed feature. (c) Ours w/o all: On the basis of (b), we remove the skip connections on $M_X, M_Y$. The experiments results on the Cityscapes val set are shown in Table 6. As we can see, adding skip connections can reduce the loss of information when the input is passed through encoders. On the other hand, the cross-network skip connections on $F_X, F_Y$ are beneficial to obtain well-trained encoder-generator pairs.

As shown in Table 7, we provide a sensitivity analysis of the hyper-parameter $\lambda$ with different values. When $\lambda$ is set a smaller value, the performance drops due to lacking $L_1$ norm loss. When $\lambda$ is set a larger value, the model lacks adversarial supervision, resulting in blurry images. Thus the hyper-parameter $\lambda$ is set 100 in our model.

5 CONCLUSION

Herein, we propose a novel framework composed of four adversarial networks, which is called dual-directional GANs, to achieve image-to-image translation. The proposed framework consists of two main sub-networks and two auxiliary sub-networks. It utilizes self-reconstruction strategy to construct auxiliary sub-networks and makes use of them to facilitate mapping learning between domains. Combining with cross-network skip connections, our model can produce more realistic images than some existing works, especially in per-pixel accuracy. Moreover, we use the proposed framework to figure out the problem where condition GAN produces sharp images containing many small, hallucinated objects. In other words, our special framework increases the use ratio of data in essence conditioned on the same dataset and obtains better encoder-generator pairs at the same time. Extensive experiments are conducted to demonstrate that our model can produce images of high perceptual quality on multiple datasets. Qualitative and quantitative comparison results suggest that dual-directional GANs is a promising method on the task of image-to-image translation.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (61673402, 61273270, and 60802069), the Natural Science Foundation of Guangdong Province (2017A030311029, 2016B010123005, 2017B090900005, and 2019B010140002), the Science and Technology Program of Guangzhou of China (201704020180 and 201604020024), and the Fundamental Research Funds for the Central Universities of China.

| Mode | Per-pixel acc. | Per-class acc. | Class IOU |
|------|----------------|----------------|-----------|
| Ours base | 0.673 | 0.233 | 0.172 |
| Ours w/o $F_X$ | 0.707 | 0.243 | 0.183 |
| Ours w/o $M_X$ | 0.715 | 0.246 | 0.185 |
| Ours(full) | 0.732 | 0.250 | 0.188 |

| Mode | Per-pixel acc. | Per-class acc. | Class IOU |
|------|----------------|----------------|-----------|
| Ours w/o all | 0.489 | 0.157 | 0.093 |
| Ours w/o $F_X, F_Y$ | 0.691 | 0.239 | 0.177 |
| Ours(full) | 0.732 | 0.250 | 0.188 |

| Values | Per-pixel acc. | Per-class acc. | Class IOU |
|--------|----------------|----------------|-----------|
| 1 | 0.655 | 0.212 | 0.162 |
| 5 | 0.691 | 0.228 | 0.171 |
| 10 | 0.710 | 0.233 | 0.179 |
| 50 | 0.734 | 0.241 | 0.182 |
| 100 | 0.732 | 0.250 | 0.188 |
| 200 | 0.726 | 0.239 | 0.181 |
| 500 | 0.708 | 0.231 | 0.175 |
REFERENCES

1. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2013). arXiv preprint arXiv:1312.6114
2. Larsen, A.B.L., et al.: Autoencoding beyond pixels using a learned similarity metric (2015). arXiv preprint arXiv:1512.09300
3. Patath, D., et al.: Context encoders: feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, pp. 2536–2544 (2016). https://doi.org/10.1109/CVPR.2016.278
4. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European Conference on Computer Vision, Springer, Cham, (2016)
5. Chen, Q., Kolotun, V.: Photographic image synthesis with cascaded refinement networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, pp. 1520–1529 (2017). https://doi.org/10.1109/ICCV.2017.168
6. Goodfellow, I.J., et al.: Generative adversarial nets. In: International Conference on Neural Information Processing Systems, pp. 2672–2680 (2014)
7. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, pp. 2650–2658 (2015). https://doi.org/10.1109/ICCV.2015.304
8. Isola, P., et al.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp. 5967–5976 (2017). https://doi.org/10.1109/CVPR.2017.632
9. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, pp. 694–711 (2016)
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp. 3431–3440 (2015). https://doi.org/10.1109/CVPR.2015.7299656
11. Ruan, C., et al.: Category-level adversaries for semantic domain adaptation. IEEE Access, 7, 83198–83208 (2019)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241 (2015). Springer
13. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: Advances in Neural Information Processing Systems, pp. 469–477 (2016)
14. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks (2015). arXiv preprint arXiv:1511.06434
15. Mirza, M., Osindero, S.: Conditional generative adversarial nets (2014). arXiv preprint arXiv:1411.1784
16. Hertzmann, A., et al.: Image analogies. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pp. 327–340 (2001)
17. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, pp. 1033–1038 vol. 2 (1999). https://doi.org/10.1109/ICCV.1999.790383
18. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, pp. 2414–2423 (2016). https://doi.org/10.1109/CVPR.2016.265
19. Wang, X., Gupta, A.: Generative image modeling using style and structure adversarial networks. In: European Conference on Computer Vision, Springer, Cham. pp. 318–335 (2016)
20. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error (2015). arXiv preprint arXiv:1511.05440
21. Yoo, D., et al.: Pixel-level domain transfer. In: European Conference on Computer Vision, Springer, Cham, pp. 517–532 (2016 October)
22. Ruan, C., Chen, D., Hu, H.: Multimodal supervised image translation. Electron. Lett. (2019)
23. Karacan, I., et al.: Learning to generate images of outdoor scenes from attributes and semantic layouts (2016). arXiv preprint arXiv:1612.00215
24. Reed, S.E., et al.: Learning what and where to draw. In: Advances in Neural Information Processing Systems, pp. 217–225 (2016)
25. Kim, T., et al.: Learning to discover cross-domain relations with generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1857–1865 (2017). JMLR.org
26. Yi, Z., et al.: Dualgan: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision, (ICCV), Venice, pp. 2686–2876 (2017). https://doi.org/10.1109/ICCV.2017.310
27. Zhu, J.Y., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, (ICCV), Venice, pp. 2242–2251 (2017). https://doi.org/10.1109/ICCV.2017.244
28. Liu, M.Y., Broeul, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems, pp. 700–708 (2017)
29. Lin, J., et al.: Conditional image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, pp. 5524–5532 (2018). https://doi.org/10.1109/CVPR.2018.00579
30. Huang, X., et al.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 172–189 (2018)
31. Kazemi, H., et al.: Unsupervised image-to-image translation using domain-specific variational information bound. In: Advances in Neural Information Processing Systems, pp. 10360–10379 (2018)
32. Lee, H.Y., et al.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 35–51 (2018)
33. Gonzalez-Garcia, A., van de Weijer, J., Bengio, Y.: Image-to-image translation for cross-domain disentanglement. In: Advances in Neural Information Processing Systems, pp. 1294–1305 (2018)
34. Zhu, J.Y., et al.: Toward multimodal image-to-image translation. In: Advances in Neural Information Processing Systems, pp. 465–476 (2017)
35. Li, C., Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks. In: European Conference on Computer Vision, pp. 702–716 (2016). Springer
36. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014). arXiv preprint arXiv:1412.6980
37. Cords, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, pp. 3213–3223 (2016). https://doi.org/10.1109/CVPR.2016.350
38. Richter, S.R., et al.: Playing for data: Ground truth from computer games. In: European Conference on Computer Vision, Nice, France, pp. 102–118 (2016). Springer
39. Tang, X., Wang, X.: Face sketch synthesis and recognition. In: Proceedings Ninth IEEE International Conference on Computer Vision, pp. 687–694 vol. 1 (2003). https://doi.org/10.1109/ICCV.2003.1238414
40. Wang, C., et al.: Perceptual adversarial networks for image-to-image transformation. IEEE Trans. Image Process. 27(8), 4066–4079 (2018)
41. Wang, Z., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13(4), 600–612 (2004)
42. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Trans. Image Process. 13(4), 600–612 (2004)
43. Zhang, H., Sindagi, V., Patel, V.M.: Image de-raining using a conditional generative adversarial network (2017). arXiv preprint arXiv:170105957

How to cite this article: Ruan C, Yuan L, Hu H, Chen D. Image translation with dual-directional generative adversarial networks. IET Comput. Vis. 2021;15:73–83. https://doi.org/10.1049/cvi.2.12011