Perceptual-DualGAN: Perceptual Losses for Image to Image Translation with Generative Adversarial Nets

Xuexin Qu[*†], Xin Wang [*‡], Zihan Wang[*†], Lei Wang[*], Lingchen Zhang[*]
[*]*Institute of Information Engineering, Chinese Academy of Sciences*
[†]*School of Cyber Security, University of Chinese Academy of Sciences*
Beijing, China
{quxuexin, wangxin, wangzihan, wanglei, zhanglingchen } @iie.ac.cn

***Abstract*—Thinking about cross-domain image-to-image translation problems, where an input image belonging to domain *U* is transformed into an output image belonging to another domain *V*. A series of typical tasks, such as style transformation, colorization, super-resolution, can be seen as cross-domain image-to-image translation tasks. Recent methods such as Conditional Generative Adversarial Networks (cGANs) make big progress in this field, but they require paired image data, which is hard to obtain. The DualGAN (Unsupervised Dual Learning for Image-to-Image Translation) architecture was proposed to solve the issue of lack of paired data. But the pixel-level reconstruction losses of DualGAN are simple. In this paper, we replace the pixel-level reconstruction losses with the perceptual reconstruction losses, and propose a more advanced framework for cross-domain image-to-image translation named perceptual-DualGAN. The perceptual reconstruction losses consist of feature reconstruction losses and style reconstruction losses, both of them are computed from pretrained loss networks. Experiments on multiple image translation tasks show that our framework almost performs superior to other methods. And the results of experiments illustrate that our framework can generate more realistic and more natural photos.**

***Index Terms*—cross-domain image-to-image translation, perceptual losses, GAN, generative model**

I. Introduction

In the filed of image processing, many typical tasks such as style transformation [16], colorization [5], [15], and super-resolution [7], [16], [18], can be regarded as image transformation tasks where a well-built framework gets some inputs of original images and outputs the desired images. If we consider the input image set as a domain *U* and the output image set as a domain *V*, we can call it cross-domain image-to-image translation.

Recently, many efficient methods for cross-domain image-to-image translation are benefiting from conditional generative adversarial networks [15], [16], [18], [28], which rely on large quantities of paired of image, such as pix2pix [15]. However, human labeling is much expensive, time-consuming, even impractical, and large quantities of data may always be unavailable. For example, there are numerous photos and sketches in nature world, but the paired photos and sketches are rare. Let us think about another example, the task of mutual conversion between daylight scenes and night scenes. In order to obtain the paired images, stationary cameras would be a good choice. However moving objects in the scene often cause varying degrees of content discrepancies.

In order to overcome the shortcoming of lacking the paired of image data, DualGAN [30] was proposed by Zili Yi *et al*. DualGAN is an unsupervised framework which combines generative adversarial nets with dual learning for cross-domain image-to-image translation. In most cases, it performs better than other methods. But DualGAN employs naive pixel-level reconstruction losses, so there are still some inadequacies. Thinking about the following case, there are two identical images offset from each other by one pixel, they would be measured different by per-pixel losses. Then we introduce the perceptual optimization. We minimize the perceptual reconstruction loss functions to train the networks rather than minimize the naive pixel-level reconstruction loss functions. In this paper, we use a pretrained loss network to define perceptual reconstruction loss functions, which means, the loss network can measure perceptual differences in content and style between images. The loss network remains fixed during the training process. We use VGG-19 [25] as loss network. The framework we proposed named perceptual-DualGAN.

In this paper, we introduce perceptual reconstruction losses. We consider perceptual reconstruction losses consist of feature reconstruction loss and style reconstruction loss, both of them can be obtained from deep convolutional neural networks [17]. As everyone knows, the reason why deep convolutional neural networks are effective in image processing tasks is that deep convolutional neural networks learn a representation of the image that makes object information increasingly explicit along processing hierarchy. Therefore, the input image is transformed into many representations that increase gradually care about the actual content of the image compared to its pixel-level values, so the style of image does.

In order to train a stable network, in this paper, we employ instance normalization [27]. Instance normalization was proposed by Dmitry *et al*. Dmitry has proved that the trained networks would be more stable by replacing batch

‡ Corresponding author.

978-1-5090-6014-6/18/$31.00 ©2018 IEEE

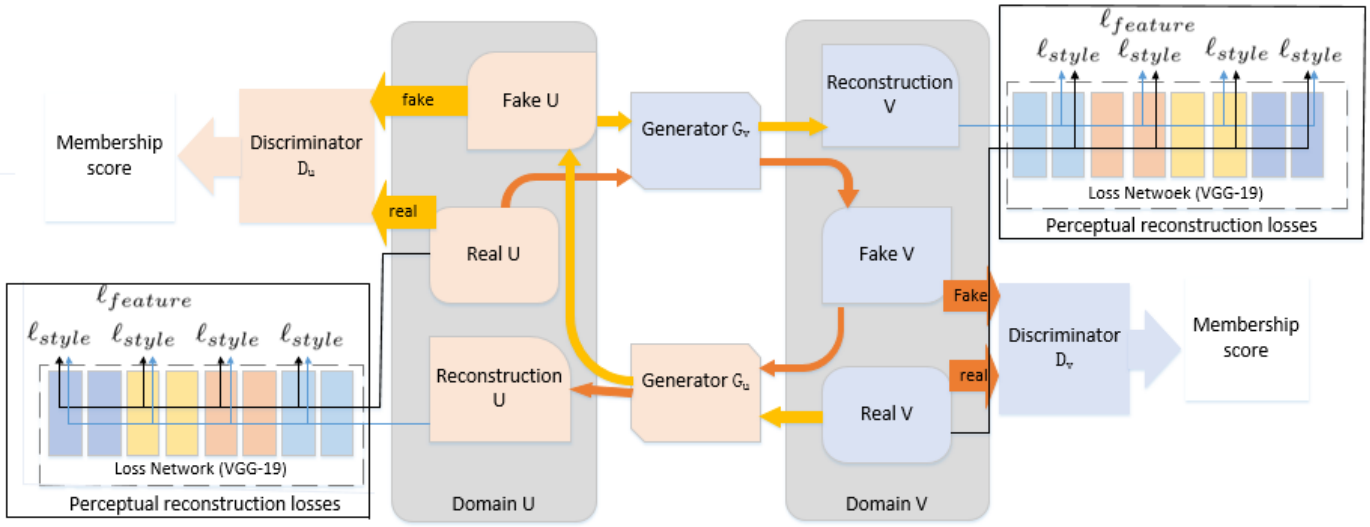

Fig. 1. Perceptual-DualGAN framework architecture and data flow chart of perceptual-DualGAN for cross-domain image-to-image translation. In this paper, we employ VGG-19 as loss network. ℓ_{style} means the style loss between the original image and the reconstruction image. $\ell_{feature}$ means the feature loss between the original image and the reconstruction image.

normalization [14] layers with instance normalization layers in training procedure, and keeping them at test time. A dramatic change in the contrast of the input image does not cause a dramatic change of the output image, when we employ instance normalization rather than batch normalization. Therefore, this technique is useful for training a more stable network.

Overall, our contributions are as follows:

- On the basis of DualGAN, we propose a new method named perceptual-DualGAN for cross-domain image-to-image translation. We have made an improvement in reconstruction losses on the basis of DualGAN. We replace naive pixel-level reconstruction losses with perceptual reconstruction losses. And in this paper, we use VGG-19 as loss network, which can be replaced by other models such as AlexNet [17], ResNet [13]. Therefore the perceptual-DualGAN is a flexible framework for cross-domain image-to-image translation.
- We introduce instance normalization. We illustrate that we can obtain a good performance by replacing batch normalization with instance normalization in training procedure.
- We provide both qualitative and quantitative results on image-to-image translation task using perceptual-DualGAN, showing our framework performs better than baseline models.

II. Related Work

In 2014, Since Goodfellow has been proposed the seminal paper Generative adversarial nets (GAN) [10], in a short period of time, a series of GAN-family [3], [4], [21], [22] have been proposed for a wide variety of problems. The vanilla generative adversarial nets compose of two parts: generator and discriminator. This framework corresponds to a minimax two-player game. Generator adversarial nets can learn a generator to capture the distribution of real data by introducing an adversarial discriminator that evolves to discriminate between the real data and the output of generator.

Conditional Generative Adversarial nets (cGAN) [21] was proposed by Mehdi Mirza *et al*. Then, a various variations [1], [15] of conditional generative adversarial nets have been widely applied in image generation besides cross-domain image-to-image translation. Isola *et al* investigate conditional adversarial nets [15] as general-purpose solution to image-to-image translation problems. This framework has been demonstrated to be valid on paired of data. CycleGAN [32] was proposed by Zhu *et al*, which is an unsupervised image-to-image translation framework.

Until now, one of the most powerful tools in image processing tasks is deep convolutional neural networks. Deep Convolutional Neural Networks consist of layers of small computational units that process visual information hierarchically in a feed-forward manner. Each layer of units can be understood as a collection of image filters, each of which extracts a certain feature from the input image. Thus, the output of a given layer consists of so-called feature maps: differently filtered versions of the input image. While deep convolutional neural networks are trained on image processing tasks such as image classification [17], object recognition [13]. Deep convolutional neural networks learn a representation of the image that makes object information increasingly explicit along processing hierarchy. Therefore, along the processing hierarchy of the network, the input image is transformed into quite a few representations that increase gradually care about the actual content of the image compared to its pixel-level values. In the other term, the texture information of image can be seen as the style of the image. We can also obtain a representation of the style of the image as well as the actual content of the image by employing deep convolutional neural

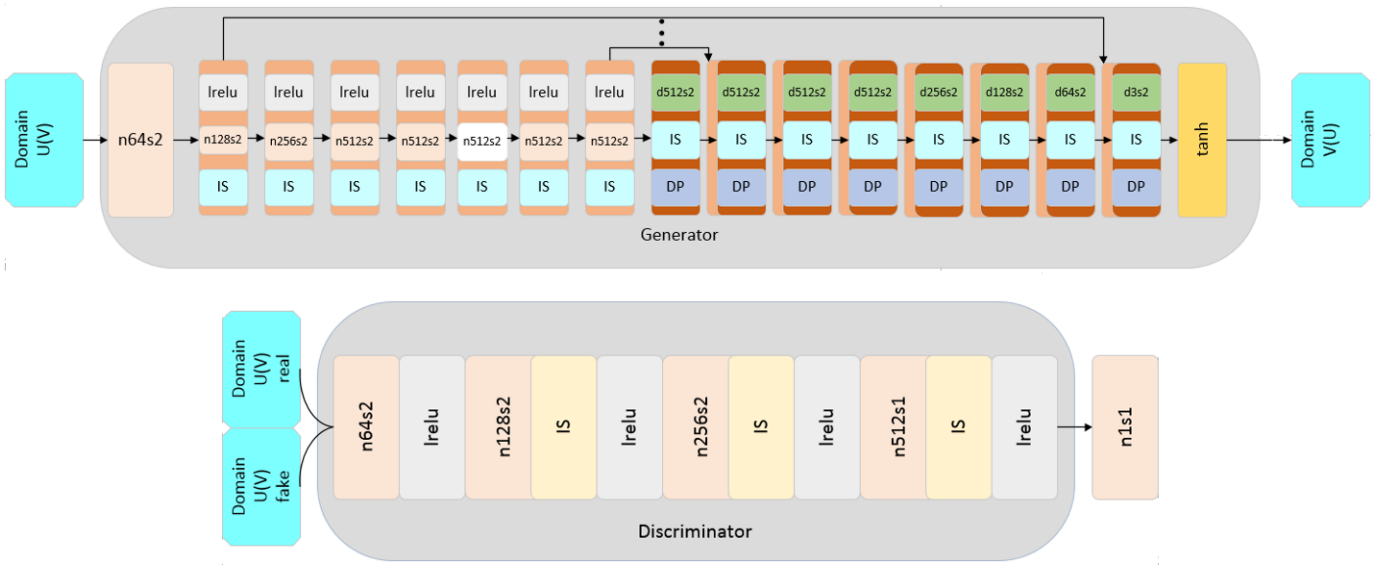

Fig. 2. An overview of generator and discriminator architecture. First row, we depict the overall generator architecture and expand the details of the generator. The generator generates domain *V* images conditioned on domain *U* images and vice versa. The discriminator discriminates images from domain *U* (domain *V*) and images from corresponding generator. A convolution with stride 2 and 64 channels is denoted as n64s2 in the figure. A deconvolution with stride 2 and 512 channels is denoted as d512s2. lrelu stands for leaky ReLU nonlinearity. IS stands for a instance normalization layer. DP stands for a dropout layer and tanh means tanh nonlinearity.

networks. Combining content loss and style loss offers us a more representative loss named perceptual loss. High-quality images can be generated by constructing and optimizing the perceptual loss based on high-level features extracted from a pretrained deep convolutional neural network such as VGG-19. Gatys *et al* achieve artistic style transfer [9] by minimizing the feature reconstruction loss, which combines the content of one image with the style of another image. In this paper, the effective of perceptual loss has also been demonstrated.

Dual learning [12] was first proposed by Xia *et al*, which was applied in machine translation to solve the problem of shortage of labeled data. The key idea of dual learning is to set up a dual-learning game which involves two agents, each of whom only understands one language. Dual-learning game can evaluate how likely the translated sentences are natural sentences in targeted language by utilizing the feedback signals that were generated by two agents.

Zili Yi *et al* proposed DualGAN which combined dual learning with GAN. DualGAN is an unsupervised framework for cross-domain image-to-image translation. However, because DualGAN use the simple pixel-level reconstruction losses, there will be a variety of problems. For example, considering two identical images offset from each other by one pixel, despite their perceptual similarity, they would be very different as measured by per-pixel losses. To overcome this problem, we introduce perceptual loss functions into GAN, named perceptual-DualGAN. This framework can learn a much better translator for cross-domain image-to-image translation on the basis of DualGAN by replacing pixel-level reconstruction losses with perceptual losses.

III. METHOD

In this paper, we propose perceptual-DualGAN framework. This framework employ perceptual reconstruction losses. And in order to make the training procedure stable, we employ Improved Training of Wasserstein GAN (WGAN-GP) [11] and replace batch normalization with instance normalization in the training procedure. Experimental results show that our framework is superior to baseline models.

First, let us look back on what is cross-domain image-to-image translation. Assuming there are two sets of unpaired and unlabeled images sample from domain U and domain V respectively. The cross-domain image-to-image translation means to learn a mapping from images belonging to the domain U to images belonging to the domain V and vice versa. As is known to all, generative adversarial nets are one of the most popular generative models. Therefore we employ generative adversarial nets to solve cross-domain image-to-image translation problem.

In our framework, there are a pair of GANs. The generator G_v maps an image u ($u \in U$) to an image v ($v \in V$). There is the generator G_u maps an image v ($v \in V$) to an image u ($u \in U$). Besides, there are not only two generators G_v and G_u but also two discriminators D_v and D_u. The discriminator D_v discriminates between the fake outputs of G_v and the real members of domain V. Analogously, the discriminator D_u discriminates between the fake outputs of G_u and the real members of domain U. More details about framework are illustrated in Fig.1.

As show in Fig.1, the image u belongs to domain U is translated to domain V by G_v. The discriminator D_v evaluates that how well the translation $G_v(u)$ fits in domain V. Then,

the $G_v(u)$ is translated back to domain U by using G_u, the output $G_u(G_v(u))$ can be seen as reconstructed version of u. Analogously, the image v belongs to domain V is translated to domain U by G_u. The discriminator D_u evaluates that how well the translation $G_u(v)$ fits in domain U. Then, the $G_u(v)$ is translated back to domain V by using G_v, the output $G_v(G_u(v))$ can be seen as reconstructed version of v. In order to train perceptual-DualGAN well, the generators G_v and G_u are optimized to output fake images to blind D_v and D_u respectively, and to minimize the reconstruction losses.

A. Objective

Generative adversarial nets are appealing generative models that cast generative modeling as a minimax game between a generator network and a discriminator network. A discriminator network discriminates between the generator's output and true data. However, Generative adversarial nets suffer from training instability and model collapse [2]. In this paper, we employ the loss format advocated by WGAN-GP rather than the method advocated by Wasserstein GAN (WGAN) [2] and the sigmoid cross-entropy loss adopted in vanilla GAN [10]. It has been demonstrated that WGAN-GP performs better than standard WGAN and enables stable training of a wide variety of GAN architectures with almost no hyperparameter tuning. Therefore, the corresponding loss functions used in D_u and D_v are defined as:

$$\begin{aligned} L_u^d = &\mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}[D_u(\tilde{x})] - \mathop{\mathbb{E}}_{x\sim\mathbb{P}_u}[D_u(x)] \\ &+\lambda \mathop{\mathbb{E}}_{\bar{x}\sim\mathbb{P}_{\bar{x}}}[(\| \bigtriangledown_{\bar{x}} D_u(\bar{x})\|_2 - 1)^2] \end{aligned} \tag{1}$$

$$\begin{aligned} L_v^d = &\mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}[D_v(\tilde{x})] - \mathop{\mathbb{E}}_{x\sim\mathbb{P}_v}[D_v(x)] \\ &+\lambda \mathop{\mathbb{E}}_{\bar{x}\sim\mathbb{P}_{\bar{x}}}[(\| \bigtriangledown_{\bar{x}} D_v(\bar{x})\|_2 - 1)^2] \end{aligned} \tag{2}$$

L_u^d is the objective of D_u, and L_v^d is the objective of D_v. The λ is a hyperparameter, in this paper, it is set to 10. We also implicitly define $\mathbb{P}_{\bar{x}}$ sampling uniformly along straight lines between pairs of points sampled from the data distribution ($\mathbb{P}_v$ or $\mathbb{P}_u$) and the corresponding generator distribution $\mathbb{P}_g$.

Rather than encouraging the pixel-level reconstruction losses of the reconstruction image and the original image, we instead encouraging them to have similar feature representations which are computed by the loss network Φ, in this paper we regard VGG-19 as loss network Φ. When loss network Φ processes the image x, let $\Phi_j(x)$ denote the activations of the j-th layer of the loss network Φ. As j is a convolutional layer, $\Phi_j(x)$ is a feature map of shape $K_j \times W_j \times H_j$, K_j is the feature channel in j-th layer, W_j and H_j span spatial dimensions in j-th layer. So the feature reconstruction loss can be defined as:

$$\ell_{feature}^{\Phi,j}(x,\hat{x}) = \frac{1}{K_j W_j H_j}\|\Phi_j(x) - \Phi_j(\hat{x})\|_2^2 \tag{3}$$

$\ell_{feature}^{\Phi,j}(x,\hat{x})$ means feature reconstruction loss between the original image x and the reconstruction image $\hat{x}$ in jth layer of loss network.

In order to penalize differences in style: colors, textures, common patterns, etc. We introduce style reconstruction loss [8] which was proposed by Gatys *et al*. Following the above marks, we define a matrix Ψ of shape $K_j \times W_j H_j$, then $G_j^\Phi = \Psi\Psi^T / K_j W_j H_j$. G_j^Φ is proportional to the uncentered covariance of the K_j-dimensional features, treating each grid location as an independent sample. The style reconstruction loss can be defined as:

$$\ell_{style}^{\Phi,j}(x,\hat{x}) = \|G_j^\Phi(x) - G_j^\Phi(\hat{x})\|_2^2 \tag{4}$$

$\ell_{style}^{\Phi,j}(x,\hat{x})$ means style reconstruction loss between the original image x and the reconstruction image $\hat{x}$ in jth layer of loss network.

In this paper, The objective of generators combine perceptual reconstruction losses and the loss of generator in WGAN-GP. Therefore, the objective of generators (G_u and G_v) can be defined as:

$$L_u^g = \lambda_u(\sum_j^J \ell_{style}^{\Phi,j}(u,\hat{u}) + \ell_{feature}^{\Phi,i}(u,\hat{u})) - \mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}[D_u(\tilde{x})] \tag{5}$$

$$L_v^g = \lambda_v(\sum_j^J \ell_{style}^{\Phi,j}(v,\hat{v}) + \ell_{feature}^{\Phi,i}(v,\hat{v})) - \mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}[D_v(\tilde{x})] \tag{6}$$

L_u^g is the objective of G_u, and L_v^g is the objective of G_v. We successively select the 1-th, 3-th, 5-th, 9-th, 13-th leaky ReLU nonlinearity layer as j and 10-th leaky ReLU nonlinearity layer as i, λ_u and λ_v are hyperparameters, both of them are set to 20. And $\mathbb{P}_g$ refers to corresponding data distribution from generator (G_u or G_v).

Finally, in order to encourage spatial smoothness in the output images, we also introduce total variation regularization [6], [20].

B. Network Configuration

An intuitive idea is that the outputs of methods for cross-domain image-to-image translation should not depend on the contrast of the input image, in other words, when the framework gets some inputs, if the contrast of inputs are different, the outputs should not change dramatically. Thus, the generator should discard contrast information of the inputs. Therefore, we introduce instance normalization also named contrast normalization. In this paper, $x \in R^{T\times W\times H\times K}$ is an input tensor containing a batch of T images, x_{twhk} denote the $twhk$-th element, where t is the index of image in the batch, w and h span spatial dimensions, k is the feature channel (color channel). Then, there is a simple version of instance normalization:

$$y_{twhk} = \frac{x_{twhk}}{\sum_{w=1}^{W}\sum_{h=1}^{H} x_{twhk}} \tag{7}$$

Let us review batch normalization:

$$
\begin{aligned}
y_{twhk} &= \frac{x_{twhk} - \mu_k}{\sqrt{\sigma_k^2 + \epsilon}} \\
\mu_k &= \frac{1}{HWT} \sum_{t=1}^{T} \sum_{w=1}^{W} \sum_{h=1}^{H} x_{twhk} \\
\sigma_k^2 &= \frac{1}{HWT} \sum_{t=1}^{T} \sum_{w=1}^{W} \sum_{h=1}^{H} (x_{twhk} - \mu_k)^2
\end{aligned} \tag{8}
$$

Because generator always uses convolution, downsampling, upsampling and batch normalization, it may be difficult to learn a highly nonlinear contrast normalization function as a combination of such layers. Furthermore, in order to combine the effects of batch normalization and instance normalization, we employ the advanced instance normalization layer:

$$
\begin{aligned}
y_{twhk} &= \frac{x_{twhk} - \mu_{tk}}{\sqrt{\sigma_{tk}^2 + \epsilon}} \\
\mu_{tk} &= \frac{1}{HW} \sum_{w=1}^{W} \sum_{h=1}^{H} x_{twhk} \\
\sigma_{tk}^2 &= \frac{1}{HW} \sum_{w=1}^{W} \sum_{h=1}^{H} (x_{twhk} - \mu_{tk})^2
\end{aligned} \tag{9}
$$

In perceptual-DualGAN framework, we replace batch normalization with advanced instance normalization everywhere in the generator network, so that instance-sepcific mean and covariance mean shift problem can be solved in training procedure.

Perceptual-DualGAN consists of two generators G_u and G_v, two discriminators D_u and D_v. Two generators are constructed with identical layers, so discriminators do. The generator and discriminator architecture are showed in Fig.2. In perceptual-DualGAN framework, the generators employ "U-Net" [23] architecture, which is an encoder-decoder with skip connections between mirrored layers in the encoder and decoder stacks. For many image translation tasks, it is very important to preserve low-level information of image in the whole process. For example, in the case of image colorization, the input and output share the location of prominent edges. Without the skip layers, the input has to be passed through a series of layers, until a bottleneck layer. In that way, the output will lose a lot of information. On the basis of "U-Net", we configure the generator with equal number of downsampling, upsampling and instance normalization layers. More details in generator are illustrated in Fig.2.

As for discriminators, we introduce the Markovian Patch-GAN [19] architecture, an efficient method for training generative neural networks for texture synthesis. With the Markovian Patch-GAN architecture, the discriminator models high-frequencies of image by penalizing structure at the scale of patches. Specifically, in our implementation, we just run this discriminator convolutionally across the image as usual, but, the output of the discriminator is not one scalar but a patch of scores. The output of discriminator means the discriminator tries to classify each patch in an image is real or fake. Averaging a patch of scores to provide the ultimate score named membership score. More details about discriminator are also showed in Fig.2. It has also been demonstrated that the Markovian Patch-GAN requires fewer parameters, runs faster.

C. Training Procedure

After we have built the network, the most important thing is that how to train the network efficiently and stably. As everyone knows, generative adversarial nets are notoriously hard to train, it is a serious problem until the appearance of WGAN-GP, which was proposed by Ishaan *et al* and was widely used because of its stability and effectiveness. In the perceptual-DualGAN, we train discriminators n_{critic} steps, then one step on generators. We employ mini-batch stochastic gradient descent and Adam solver. Hyperparameters α, β_1, β_2 of Adam are set to 0.0001, 0, 0.9 respectively in this paper. We typically set the number of critic iterations per generator iteration n_{critic} to 5 and assign batch size to 4. The traditional method of training GANs requires carefully balance between the generator and the discriminator. Because while discriminator improves along with training process, the generator is large probability locally saturated and results in vanishing gradients. WGAN-GP introduces Wasserstein loss which is differential almost everywhere and Gradient penalty [11] can almost solve this problem. In our training process, the

Algorithm 1 Perceptual-DualGAN training procedure. we use default values of $\lambda = 10$, $n_{critic} = 5$, $\alpha = 0.0001$, $\beta_1 = 0$, $\beta_2 = 0.9$.

Require: Image set *U*, image set *V*. A GAN u with generator parameters θ_u and discriminator with parameters ω_u. A GAN v with generator parameters θ_v and discriminator with parameters ω_v. The gradient penalty coefficient λ, the number of critic iterations per generator iteration n_{critic}, the batch size m, Adam hyperparameters α, β_1, β_2. Randomly initialize θ_i, ω_i, $i \in \{u, v\}$.

repeat
  for t = 1,...,n_{critic} **do**
    sample images$\{u^{(k)}\}_{k=1}^{m} \subseteq U$,$\{v^{(k)}\}_{k=1}^{m} \subseteq V$.
    sample a random number $\epsilon \sim U[0, 1]$.
    $\tilde{x}_v \leftarrow G_v(\{u^{(k)}\}_{k=1}^{m})$.
    $\tilde{x}_u \leftarrow G_u(\{v^{(k)}\}_{k=1}^{m})$.
    $\bar{x}_v \leftarrow \epsilon x_v + (1 - \epsilon)\tilde{x}_v$.
    $\bar{x}_u \leftarrow \epsilon x_u + (1 - \epsilon)\tilde{x}_u$.
    $L_v^d \leftarrow D_v(\tilde{x}_v) - D_v(\{v^{(k)}\}_{k=1}^{m}) + \lambda(\| \bigtriangledown_{\bar{x}_v} D_v(\bar{x}_v)\|_2 - 1)^2$.
    $L_u^d \leftarrow D_u(\tilde{x}_u) - D_u(\{u^{(k)}\}_{k=1}^{m}) + \lambda(\| \bigtriangledown_{\bar{x}_u} D_u(\bar{x}_u)\|_2 - 1)^2$.
    $\omega_v \leftarrow Adam(\bigtriangledown_{\omega_v} \frac{1}{m} L_v^d, \omega_v, \alpha, \beta_1, \beta_2)$.
    $\omega_u \leftarrow Adam(\bigtriangledown_{\omega_u} \frac{1}{m} L_u^d, \omega_u, \alpha, \beta_1, \beta_2)$.
  end for
  sample images$\{u^{(k)}\}_{k=1}^{m} \subseteq$U,$\{v^{(k)}\}_{k=1}^{m} \subseteq$V.
  $\theta_v \leftarrow Adam(\bigtriangledown_{\theta_v} \frac{-1}{m} D_v(G_v(\{u^{(k)}\}_{k=1}^{m})), \theta_v, \alpha, \beta_1, \beta_2)$
  $\theta_u \leftarrow Adam(\bigtriangledown_{\theta_u} \frac{-1}{m} D_u(G_u(\{v^{(k)}\}_{k=1}^{m})), \theta_u, \alpha, \beta_1, \beta_2)$
until convergence

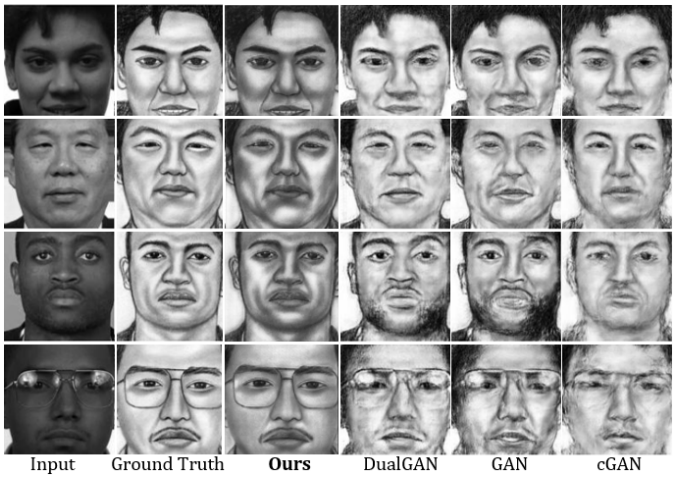

Fig. 3. Photo→Sketch translation task for faces. As shown in the illustrated, the results of our framework perceptual-DualGAN are more clear than other baseline methods.

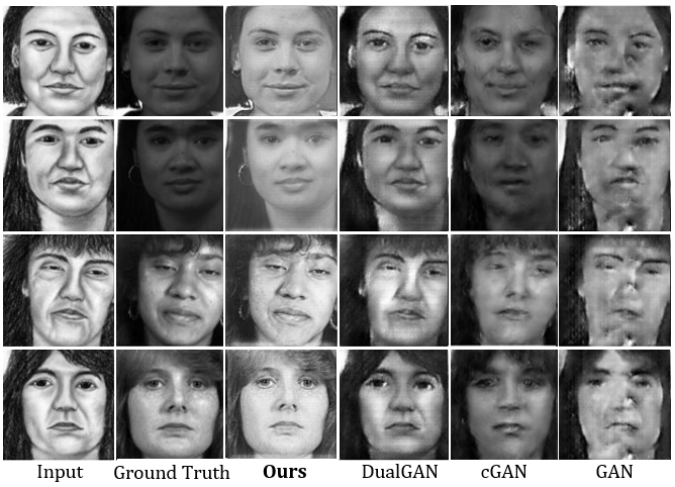

Fig. 5. Sketch→Photo translation task for faces. As we can see, our framework perceptual-DualGAN can generate more better results than baseline methods in a visually.

generators are not trained until the discriminators have been trained for n_{critic} steps. It allows the discriminators to provide more reliable gradient information.

IV. Experimental Results

In order to evaluate the capability of perceptual-DualGAN in image-to-image translation task, we conduct experiments on a variety of datasets. The results have been demonstrated that our framework is much better than other methods on most datasets.

We compare our framework perceptual-DualGAN with DualGAN, cGAN and GAN on some datasets. The paired and labeled datasets PHOTO-SKETCH [29], [31] as well as LABEL-FACADES [26] are applied to experiments. Both of datasets consist of corresponding images between two domains, and they are labeled as ground truth. The images of datasets are collected from sophisticated real world and some of them are created by artists or engineers, so the images of both datasets would not guarantee accurate feature alignment. In order to illustrate the high efficiency of our framework, we also conduct experiments on an unlabeled and unpaired dataset MATERIAL [24], which contains images of objects made of different materials, such as plastic, metal, leather, fabric and so on. Our framework can achieve material transfer by using this dataset. All of these datasets are divided two parts, training set and test set. We train models on training set and evaluate models on test set. We train the cGAN model as suggested in [15] and follow suggestion of that paper, we choose the optimal loss function and hyperparameters, more details can be seen in [15]. As for DualGAN and GAN, we train these models in an unsupervised manner. All the details including training procedure, loss function, hyperparameters and so on are followed corresponding paper [10], [30]. The way of training our model has been given in this paper, and all details can be found in the above sections of this paper. Loss functions of our framework can be found in objective section of this paper, and the configurations of our framework have been given in network configuration section of this paper. We train our model followed the training procedure section of this paper, and the value of hyperparameters have also been given in training procedure section.

Firstly, we compare our framework with DualGAN, cGAN [15] and GAN on the PHOTO-SKETCH and LABEL-

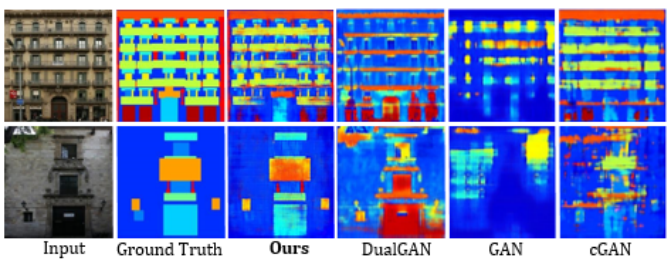

Fig. 4. Experimental results for Facades→Label translation task. The results demonstrate that our framework can retain the details such as doors, windows to maximum extent compare with other methods.

Fig. 6. Experimental results for Label→Facades translation task. The results illustrate that comparing to other models, the outputs of our framework are more vivid and more close to the ground truth.

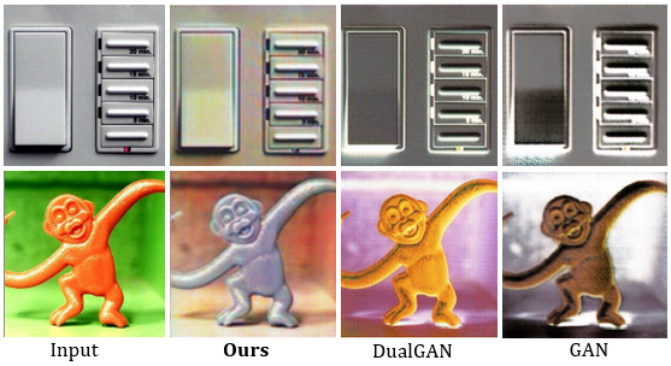

Fig. 7. Experimental results for material transfer task, plastic→metal.

FACADES datasets, the results of photo to sketch translation are illustrated in Fig.3, as well as the results of sketch to photo translation are shown in Fig.5. The results of the task of mutual conversion between facade and label are shown in Fig.4 and Fig.6 respectively. Compared to other models, the results on paired and labeled datasets illustrate that our framework can generate more realistic photo and more clear photo. The results also show that our framework can transfer image to another domain without introducing any noise. Then, because cGAN requires paired and labeled data, we just compare our framework with DualGAN and GAN on the unpaired and unlabeled dataset MATERIAL. The results of plastic to metal translation are illustrated in Fig.7, and the results of leather to fabric translation are shown in Fig.8. The results on unpaired and unlabeled dataset also show that our framework can generate less blurry photos and more natural photos. In a word, in almost cases, the results of our framework contain fewer artifacts and are more close to the ground truth on the labeled and paired dataset, and look more vivid and natural on all of datasets. On the other hand, on the unlabeled and unpaired dataset MATERIAL, the results reveal that our framework can retain some characteristics of inputs which is beyond the scope of transfer task to the maximum extent, such as color. In order to be convincing, all experiments are conducted under the same situation where the computer equips with a single GTX 1080 GPU.

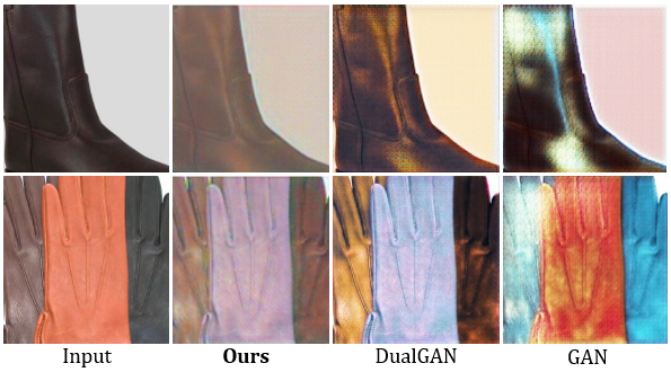

Fig. 8. Experimental results for material transfer task, leather→fabric.

V. Conclusion

In this paper, on the basis of DualGAN, we introduce perceptual reconstruction losses as well as instance normalization. The perceptual reconstruction losses consist of feature reconstruction losses and style reconstruction losses, both of which can be obtained from pretrained deep neural network. Then, we propose a more advanced and flexible framework named perceptual-DualGAN, an efficient and novel unsupervised dual learning framework for image-to-image translation. Due to the unsupervised characteristic of perceptual-DualGAN, it can be applied to numerous real word applications. In this paper, the results have demonstrated that our framework preforms better than other methods such as traditional GAN, conditional GAN. results also show that our framework is a promising method for image-to-image translation tasks.

Acknowledgment

This work was partially supported by National Key R&D Plan No.2016QY02D0401, and National Natural Science Foundation of China No.U163620068.

References

[1] M. E. Abbasnejad, Q. Shi, I. Abbasnejad, A. v. d. Hengel, and A. Dick. Bayesian conditional generative adverserial networks. *arXiv preprint arXiv:1706.05477*, 2017.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[3] D. Berthelot, T. Schumm, and L. Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[4] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

[5] Z. Cheng, Q. Yang, and B. Sheng. Deep colorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 415–423, 2015.

[6] E. d'Angelo, A. Alahi, and P. Vandergheynst. Beyond bits: Reconstructing images from local binary descriptors. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 935–938. IEEE, 2012.

[7] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016.

[8] L. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 262–270, 2015.

[9] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

[12] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T. Liu, and W.-Y. Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828, 2016.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[16] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.
[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
[18] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.
[19] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016.
[20] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.
[21] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
[22] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
[23] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.
[24] L. Sharan, R. Rosenholtz, and E. Adelson. Material perception: What can you see in a brief glance? *Journal of Vision*, 9(8):784–784, 2009.
[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[26] R. Tyleček and R. Šára. Spatial pattern templates for recognition of objects with regular structure. In *German Conference on Pattern Recognition*, pages 364–374. Springer, 2013.
[27] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.
[28] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*, pages 318–335. Springer, 2016.
[29] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2009.
[30] Z. Yi, H. Zhang, P. T. Gong, et al. Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv preprint arXiv:1704.02510*, 2017.
[31] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 513–520. IEEE, 2011.
[32] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.