Conditional Image-to-Image translation

Jianxin Lin[1] Yingce Xia[1] Tao Qin[2] Zhibo Chen[1] Tie-Yan Liu[2]
[1]University of Science and Technology of China [2]Microsoft Research Asia
linjx@mail.ustc.edu.cn yingce.xia@gmail.com
{taoqin, tie-yan.liu}@microsoft.com chenzhibo@ustc.edu.cn

Abstract

Image-to-image translation tasks have been widely investigated with Generative Adversarial Networks (GANs) and dual learning. However, existing models lack the ability to control the translated results in the target domain and their results usually lack of diversity in the sense that a fixed image usually leads to (almost) deterministic translation result. In this paper, we study a new problem, conditional image-to-image translation, which is to translate an image from the source domain to the target domain conditioned on a given image in the target domain. It requires that the generated image should inherit some domain-specific features of the conditional image from the target domain. Therefore, changing the conditional image in the target domain will lead to diverse translation results for a fixed input image from the source domain, and therefore the conditional input image helps to control the translation results. We tackle this problem with unpaired data based on GANs and dual learning. We twist two conditional translation models (one translation from A domain to B domain, and the other one from B domain to A domain) together for inputs combination and reconstruction while preserving domain independent features. We carry out experiments on men's faces from-to women's faces translation and edges to shoes&bags translations. The results demonstrate the effectiveness of our proposed method.

1. Introduction

Image-to-image translation covers a large variety of computer vision problems, including image stylization [4], segmentation [13] and saliency detection [5]. It aims at learning 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 example, in the aforementioned three tasks, an input image might be converted to a portrait similar to Van Gogh's styles, a heat map splitted into different regions, or a pencil sketch, while the edges and outlines remain unchanged. Since it is usually hard to collect a large amount of parallel data for such tasks, unsupervised learning algorithms have been widely adopted. Particularly, the generative adversarial networks (GAN) [6] and dual learning [7, 21] are extensively studied in image-to-image translations. [22, 9, 25] tackle image-to-image translation by the aforementioned two techniques, where the GANs are used to ensure the generated images belonging to the target domain, and dual learning can help improve image qualities by minimizing reconstruction loss.

An implicit assumption of image-to-image translation is that an image contains two kinds of features[1]: *domain-independent features*, which are preserved during the translation (i.e., the edges of face, eyes, nose and mouse while translating a man' face to a woman' face), and *domain-specific features*, which are changed during the translation (i.e., the color and style of the hair for face image translation). Image-to-Image translation aims at transferring images from the source domain to the target domain by preserving domain-independent features while replacing domain-specific features.

While it is not difficult for existing image-to-image translation methods to convert an image from a source domain to a target domain, it is not easy for them to control or manipulate the style in fine granularity of the generated image in the target domain. Consider the gender transform problem studied in [9], which is to translate a man's photo to a woman's. Can we translate Hillary's photo to a man' photo with the hair style and color of Trump? DiscoGAN [9] can indeed output a woman's photo given a man's photo as input, but cannot control the hair style or color of the output image. DualGAN [22, 25] cannot implement this kind of fine-granularity control neither. To fulfill such a blank in image translation, we propose the concept of *conditional image-to-image* translation, which can specify domain-specific features in the target domain, carried by another input image from the target domain. An example of conditional image-to-image translation is shown in Figure 1, in

[1] Note that the two kinds of features are relative concepts, and domain-specific features in one task might be domain-independent features in another task, depending on what domains one focuses on in the task.

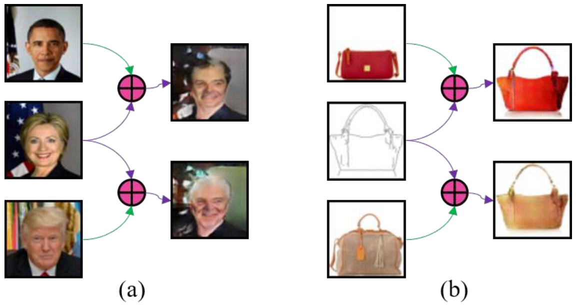

Figure 1. Conditional image-to-image translation. (a) Conditional women-to-men photo translation. (b) Conditional edges-to-handbags translation. The purple arrow represents translation flow and the green arrow represents the conditional information flow.

which we want to convert Hillary's photo to a man's photo. As shown in the figure, with an addition man's photo as input, we can control the translated image (e.g., the hair color and style).

1.1. Problem Setup

We first define some notations. Suppose there are two image domains $\mathcal{D}_A$ and $\mathcal{D}_B$. Following the implicit assumption, an image $x_A \in \mathcal{D}_A$ can be represented as $x_A = x_A^i \oplus x_A^s$, where x_A^i's are domain-independent features, x_A^s's are domain-specific features, and $\oplus$ is the operator that can merge the two kinds of features into a complete image. Similarly, for an image $x_B \in \mathcal{D}_B$, we have $x_B = x_B^i \oplus x_B^s$. Take the images in Figure 1 as examples: (1) If the two domains are man's and woman's photos, the domain-independent features are individual facial organs like eyes and mouths and the domain-specific features are beard and hair style. (2) If the two domains are real bags and the edges of bags, the domain-independent features are exactly the edges of bags themselves, and the domain-specific are the colors and textures.

The problem of conditional image-to-image translation from domain $\mathcal{D}_A$ to $\mathcal{D}_B$ is as follows: Taken an image $x_A \in \mathcal{D}_A$ as input and an image $x_B \in \mathcal{D}_B$ as conditional input, outputs an image x_{AB} in domain $\mathcal{D}_B$ that keeping the domain-independent features of x_A and combining the domain-specific features carried in x_B, i.e.,

$$x_{AB} = G_{A\to B}(x_A, x_B) = x_A^i \oplus x_B^s, \tag{1}$$

where $G_{A\to B}$ denotes the translation function. Similarly, we have the reverse conditional translation

$$x_{BA} = G_{B\to A}(x_B, x_A) = x_B^i \oplus x_A^s. \tag{2}$$

For simplicity, we call $G_{A\to B}$ the forward translation and $G_{B\to A}$ the reverse translation. In this work we study how to learn such two translations.

1.2. Our Results

There are three main challenges in solving the conditional image translation problem. The first one is how to extract the domain-independent and domain-specific features for a given image. The second is how to merge the features from two different domains into a natural image in the target domain. The third one is that there is no parallel data for us to learn such the mappings.

To tackle these challenges, we propose the *conditional dual-GAN* (briefly, cd-GAN), which can leverage the strengths of both GAN and dual learning. Under such a framework, the mappings of two directions, $G_{A\to B}$ and $G_{B\to A}$, are jointly learned. The model of cd-GAN follows the encoder-decoder based framework: the encoder is used to extract the domain-independent and domain-specific features and the decoder is to merge the two kinds of features to generate images. We chose GAN and dual learning due to the following considerations: (1) The dual learning framework can help learn to extract and merge the domain-specific and domain-independent features by minimizing carefully designed reconstruction errors, including reconstruction errors of the whole image, the domain-independent features, and the domain-specific features. (2) GAN can ensure that the generated images well mimic the natural images in the target domain. (3) Both dual learning [7, 22, 25] and GAN [6, 19, 1] work well under unsupervised settings.

We carry out experiments on different tasks, including face-to-face translation, edge-to-shoe translation, and edge-to-handbag translation. The results demonstrate that our network can effectively translate image with conditional information and robust to various applications.

Our main contributions lie in two folds: (1) We define a new problem, conditional image-to-image translation, which is a more general framework than conventional image translation. (2) We propose the cd-GAN algorithm to solve the problem in an end-to-end way.

The remaining parts are organized follows. We introduce related work in Section 2 and present the details of cd-GAN in Section 3, including network architecture and the training algorithm. Then we report experimental results in Section 4 and conclude in Section 5.

2. Related Work

Image generation has been widely explored in recent years. Models based on variational autoencoder (VAE) [11] aim to improve the quality and efficiency of image generation by learning an inference network. GANs [6] were firstly proposed to generate images from random variables by a two-player minimax game. Researchers have been exploited the capability of GANs for various image generation tasks. [1] proposed to synthesize images at multiple resolutions with a Laplacian pyramid of adversarial generators and discriminators, and can condition on class labels for controllable generation. [19] introduced a class of deep convolutional generative networks (DCGANs) for high-quality image generation and unsupervised image classification tasks.

Instead of learning to generate image samples from scratch (i.e., random vectors), the basic idea of image-to-image translation is to learn a parametric translation function that transforms an input image in a source domain to an image in a target domain. [13] proposed a fully convolutional network (FCN) for image-to-segmentation translation. Pix2pix [8] extended the basic FCN framework to other image-to-image translation tasks, including label-to-street scene and aerial-to-map. Meanwhile, pix2pix utilized adversarial training technique to ensure high-level domain similarity of the translation results.

The image-to-image models mentioned above require paired training data between the source and target domains. There is another line of works studying unpaired domain translation. Based on adversarial training, [3] and [2] proposed algorithms to jointly learn to map latent space to data space and project the data space back to latent space. [20] presented a domain transfer network (DTN) for unsupervised cross-domain image generation employing a compound loss function including multiclass adversarial loss and f-constancy component, which could generate convincing novel images of previously unseen entities and preserve their identity. [7] developed a dual learning mechanism which can enable a neural machine translation system to automatically learn from unlabeled data through a dual learning game. Following the idea of dual learning, DualGAN [22], DiscoGAN [9] and CycleGAN [25] were proposed to tackle the unpaired image translation problem by training two cross domain transfer GANs at the same time. [15] proposed to utilize dual learning for semantic image segmentation. [14] further proposed a conditional CycleGAN for face super-resolution by adding facial attributes obtained from human annotation. However, collecting a large amount of such human annotated data can be hard and expensive.

In this work, we study a new setting of image-to-image translation, in which we hope to control the generated images in fine granularity with unpaired data. We call such a new problem *conditional image-to-image translation*.

3. Conditional Dual GAN

Figure 2 shows the overall architecture of the proposed model, in which the left part is an encoder-decoder based framework for image translation and the right part includes additional components introduced to train the encoder and decoder.

3.1. The Encoder-Decoder Framework

As shown in the figure, there are two encoders e_A and e_B and two decoders g_A and g_B.

The encoders serve as feature extractors, which take an image as input and output the two kinds of features, domain-independent features and domain-specific features, with the corresponding modules in the encoders. In particular, given two images x_A and x_B, we have

$$(x_A^i, x_A^s) = e_A(x_A); \quad (x_B^i, x_B^s) = e_B(x_B). \tag{3}$$

If only looking at the encoder, there is no difference between the two kinds of features. It is the remaining parts of the overall model and the training process that differentiate the two kinds of features. More details are discussed in Section 3.3.

The decoders serve as generators, which take as inputs the domain-independent features from the image in the source domain and the domain-specific features from the image in the target domain and output a generated image in the target domain. That is,

$$x_{AB} = g_B(x_A^i, x_B^s); \quad x_{BA} = g_A(x_B^i, x_A^s). \tag{4}$$

3.2. Training Algorithm

We leverage dual learning techniques and the GAN techniques to train the encoders and decoders. The optimization process is shown in the right part of Figure 2.

3.2.1 GAN loss

To ensure the generated x_{AB} and x_{BA} are in the corresponding domains, we employ two discriminators d_A and d_B to differentiate the real images and synthetic ones. d_A (or d_B) takes an image as input and outputs a probability indicating how likely the input is a natural image from domain $\mathcal{D}_A$ (or $\mathcal{D}_B$). The objective function is

$$\begin{aligned}\ell_{\text{GAN}} =& \log(d_A(x_A)) + \log(1 - d_A(x_{BA})) \\ &+ \log(d_B(x_B)) + \log(1 - d_B(x_{AB})).\end{aligned} \tag{5}$$

The goal of the encoders and decoders e_A, e_B, g_A, g_B is to generate images as similar to natural images and fool the discriminators d_A and d_B, i.e., they try to minimize ℓ_{GAN}. The goal of d_A and d_B is to differentiate generated images from natural images, i.e., they try to maximize ℓ_{GAN}.

3.2.2 Dual learning loss

The key idea of dual learning is to improve the performance of a model by minimizing the reconstruction error.

To reconstruct the two images $\hat{x}_A$ and $\hat{x}_B$, as shown in Figure 2, we first extract the two kinds of features of the generated images:

$$(\hat{x}_A^i, \hat{x}_B^s) = e_B(x_{AB}); \quad (\hat{x}_B^i, \hat{x}_A^s) = e_A(x_{BA}), \tag{6}$$

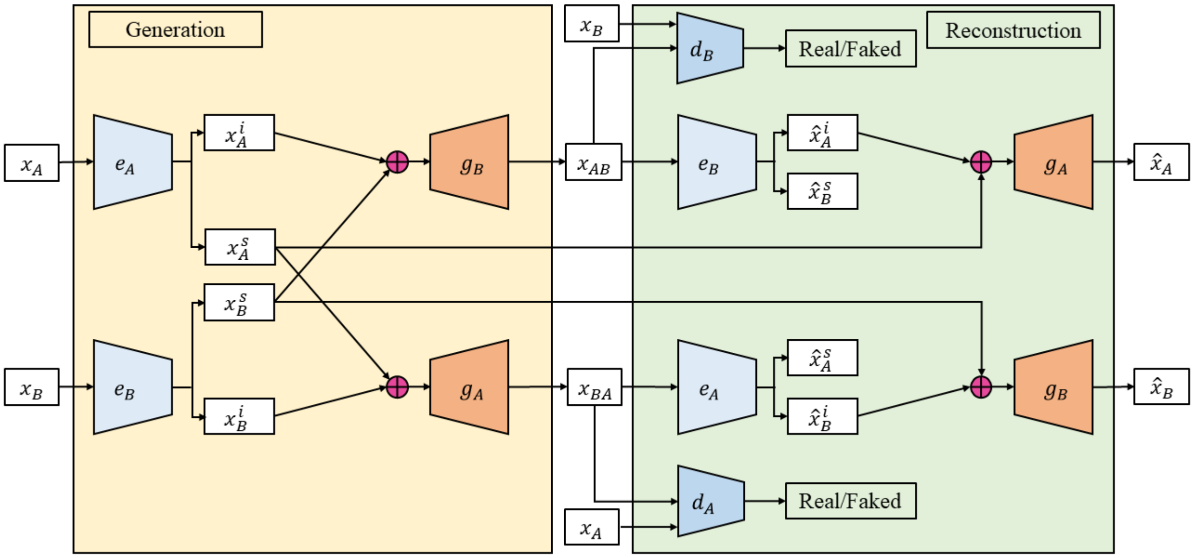

Figure 2. Architecture of the proposed conditional dual GAN (cd-GAN).

and then reconstruct images as follows:

$$\hat{x}_A = g_A(\hat{x}_A^i, x_A^s); \quad \hat{x}_B = g_B(\hat{x}_B^i, x_B^s). \tag{7}$$

We evaluate the reconstruction quality from three aspects: the image level reconstruction error $\ell_{\text{dual}}^{\text{im}}$, the reconstruction error $\ell_{\text{dual}}^{\text{di}}$ of the domain-independent features, and the reconstruction error $\ell_{\text{dual}}^{\text{ds}}$ of the domain-specific features as follows:

$$\ell_{\text{dual}}^{\text{im}}(x_A, x_B) = \|x_A - \hat{x}_A\|^2 + \|x_B - \hat{x}_B\|^2, \tag{8}$$

$$\ell_{\text{dual}}^{\text{di}}(x_A, x_B) = \|x_A^i - \hat{x}_A^i\|^2 + \|x_B^i - \hat{x}_B^i\|^2, \tag{9}$$

$$\ell_{\text{dual}}^{\text{ds}}(x_A, x_B) = \|x_A^s - \hat{x}_A^s\|^2 + \|x_B^s - \hat{x}_B^s\|^2. \tag{10}$$

Compared with the existing dual learning approaches [22] which only consider the image level reconstruction error, our method considers more aspects and therefore is expected to achieve better accuracy.

3.2.3 Overall training process

Since the discriminators only impact the GAN loss ℓ_{GAN}, we only use this loss to compute the gradients and update d_A and d_B. In contrast, the encoders and decoders impact all the 4 losses (i.e., the GAN loss and three reconstruction errors), we use all the 4 objectives to compute gradients and update models for them. Note that since the 4 objectives are of different magnitudes, their gradients may vary a lot in terms of magnitudes. To smooth the training process, we normalize the gradients so that their magnitudes are comparable across 4 losses. We summarize the training process in Algorithm 1.

Algorithm 1 cd-GAN training process

```
Require: Training images $\{x_{A,i}\}_{i=1}^m \subset \mathcal{D}_A$, $\{x_{B,j}\}_{j=1}^m \subset \mathcal{D}_B$, batch size $K$, optimizer $Opt(\cdot,\cdot)$;
1: Randomly initialize $e_A$, $e_B$, $g_A$, $g_B$, $d_A$ and $d_B$.
2: Randomly sample a minibatch of images and prepare the data pairs $\mathcal{S} = \{(x_{A,k}, x_{B,k})\}_{k=1}^K$.
3: For any data pair $(x_{A,k}, x_{B,k}) \in \mathcal{S}$, generate conditional translations by Eqn.(3,4), and reconstruct the images by Eqn.(6,7);
4: Update the discriminators as follows:
     $d_A \leftarrow Opt(d_A, (1/K)\nabla_{d_A}\sum_{k=1}^K \ell_{\text{GAN}}(x_{A,k}, x_{B,k}))$,
     $d_B \leftarrow Opt(d_B, (1/K)\nabla_{d_B}\sum_{k=1}^K \ell_{\text{GAN}}(x_{A,k}, x_{B,k}))$;
5: For each $\Theta \in \{e_A, e_B, g_A, g_B\}$, compute the gradients
     $\Delta_{\text{GAN}} = (1/K)\nabla_\Theta \sum_{k=1}^K \ell_{\text{GAN}}(x_{A,k}, x_{B,k})$,
     $\Delta_{\text{im}} = (1/K)\nabla_\Theta \sum_{k=1}^K \ell_{\text{dual}}^{\text{im}}(x_{A,k}, x_{B,k})$,
     $\Delta_{\text{di}} = (1/K)\nabla_\Theta \sum_{k=1}^K \ell_{\text{dual}}^{\text{di}}(x_{A,k}, x_{B,k})$,
     $\Delta_{\text{ds}} = (1/K)\nabla_\Theta \sum_{k=1}^K \ell_{\text{dual}}^{\text{ds}}(x_{A,k}, x_{B,k})$,
     normalize the four gradients to make their magnitudes comparable, sum them to obtain $\Delta$, and $\Theta \rightarrow Opt(\Theta, \Delta)$.
6: Repeat step 2 to step 6 until convergence
```

In Algorithm 1, the choice of optimizers $Opt(\cdot,\cdot)$ is quite flexible, whose two inputs are the parameters to be optimized and the corresponding gradients. One can choose different optimizers (e.g. Adam [10], or nesterov gradient descend [18]) for different tasks, depending on common practice for specific tasks and personal preferences. Besides, the e_A, e_B, g_A, g_B, d_A, d_B might refer to either the models themselves, or their parameters, depending on the context.

3.3. Discussions

Our proposed framework can learn to separate the domain-independent features and domain-specific features. In Figure 2, consider the path of $x_A \to e_A \to x_A^i \to g_B \to x_{AB}$. Note that after training we ensure that x_{AB} is an image in domain $\mathcal{D}_B$ and the features x_A^i are still preserved in x_{AB}. Thus, x_A^i should try to inherent the features that are independent to domain $\mathcal{D}_A$. Given that x_A^i is domain independent, it is x_B^s that carries information about domain $\mathcal{D}_B$. Thus, x_B^s is domain-specific features. Similarly, we can see that x_A^s is domain-specific and x_B^i is domain-independent.

DualGAN [22], DiscoGAN [9] and CycleGAN [25] can be treated as simplified versions of our cd-GAN, by removing the domain-specific features. For example, in CycleGAN, given an $x_A \in \mathcal{D}_A$, any $x_{AB} \in \mathcal{D}_B$ is a legal translation, no matter what $x_B \in \mathcal{D}_B$ is. In our work, we require that the generated images should match the inputs from two domains, which is more difficult.

Furthermore, cd-GAN works for both symmetric translations and asymmetric translations. In symmetric translations, both directions of translations need conditional inputs (illustrated in Figure 1(a)). In asymmetric translations, only one direction of translation needs a conditional image as input (illustrated in Figure 1(b)). That is, the translation from bag to edge does not need another edge image as input; even given an additional edge image as the conditional input, it does not change or help to control the translation result.

For asymmetric translations, we only need to slightly modify objectives for cd-GAN training. Suppose the translation direction of $G_{B\to A}$ does not need conditional input. Then we do not need to reconstruct the domain-specific features x_A^s. Accordingly, we modify the error of domain-specific features as follows, and other 3 losses do not change.

$$\ell_{\text{dual}}^{\text{ds}}(x_A, x_B) = \|x_B^s - \hat{x}_B^s\|^2 \tag{11}$$

4. Experiments

We conduct a set of experiments to test the proposed model. We first describe experimental settings, and then report results for both symmetric translations and asymmetric translations. Finally we study individual components and loss functions of the proposed model.

4.1. Settings

For all experiments, the networks take images of 64×64 resolution as inputs. The encoders e_A and e_B start with 3 convolutional layers, each convolutional layer followed by leaky rectified linear units (Leaky ReLU) [16]. Then the network is splitted into two branches: in one branch, a convolutional layer is attached to extract domain-independent features; in the other branch, two fully-connected layers are attached to extract domain-specific features. Decoder networks g_A and g_B contain 4 deconvolutional layers with ReLU units [17], except for the last layer using $\tanh$ activation function. The discriminators d_A and d_B consist of 4 convolution layers, two fully-connected layers. Each layer is followed by Leaky ReLU units except for the last layer using sigmoid activation function. Details (e.g., number and size of filters, number of nodes in fully-connected layers) can be found in the supplementary document.

We use Adam [10] as the optimization algorithm with learning rate 0.0002. Batch normalization is applied to all convolution layers and deconvolution layers except for the first and last ones. Minibatch size is fixed as 200 for all the tasks.

We implement three related baselines for comparison.

1. *DualGAN* [22, 9, 25]. DualGAN was primitively proposed for unconditional image-to-image translation which does not require conditional input. Similar to our cd-GAN, DualGAN trains two translation models jointly.

2. *DualGAN-c*. In order to enable DualGAN to utilize conditional input, we design a network as DualGAN-c. The main difference between DualGAN and DualGAN-c is that DualGAN-c translates the target outputs as Eqn.(3,4), and reconstructs inputs as $\hat{x}_A = g_A(e_B(x_{AB}))$ and $\hat{x}_B = g_B(e_A(x_{BA}))$.

3. *GAN-c*. To verify the effectiveness of dual learning, we remove the dual learning losses of cd-GAN during training and obtain GAN-c.

For symmetric translations, we carry out experiments on men-to-women face translations. We use the CelebA dataset [12], which consists of 84434 men's images (denoted as domain $\mathcal{D}_A$) and 118165 women's images (denoted as domain $\mathcal{D}_B$). We randomly choose 4732 men's images and 6379 women's images for testing, and use the rest for training. In this task, the domain-independent features are organs (e.g., eyes, nose, mouse, ? and domain-specific features refer to hair-style, beard, the usage of lipstick. For asymmetric translations, we work on edges-to-shoes and edges-to-bags translations with datasets used in [23] and [24] respectively. In these two tasks, the domain-independent features are edges and domain-specific features are colors, textures, etc.

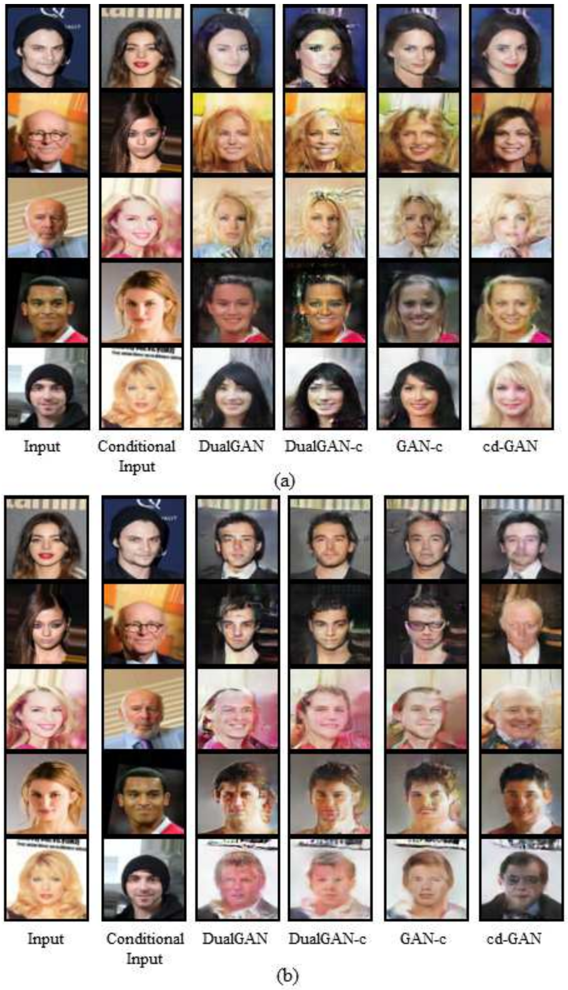

Figure 3. Conditional face-to-face translation. (a) Results of conditional men→women translation. (b) Results of conditional women→men translation.

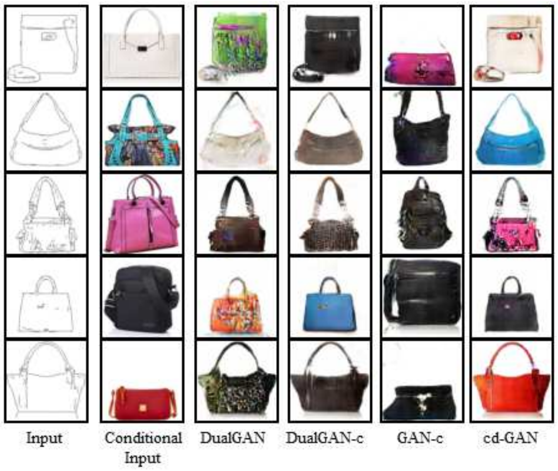

Figure 4. Results of conditional edges→handbags translation.

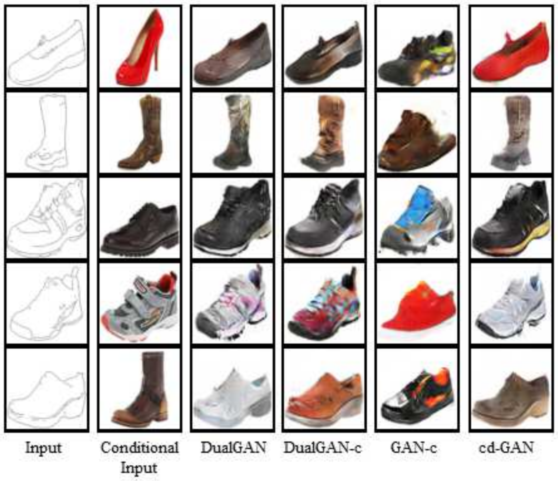

Figure 5. Results of conditional edges→shoes translation.

4.2. Results

The translation results of face-to-face, edges-to-bags and edges-to-shoes are shown in Figure 3-5 respectively.

For men-to-women translations, from Figure 3(a), we have several observations. (1) DualGAN can indeed generate woman's photo, but its results are purely based on the men's photos, since it does not take the conditional images as inputs. (2) Although taking the conditional image as input, DualGAN-c fails to integrate the information (e.g., style) from the conditional input into its translation output. (3) For GAN-c, sometimes its translation result is not relevant to the original source-domain input, e.g., the 4-th row Figure 3(a). This is because in training it is required to generate a target-domain image, but its output is not required to be similar (in certain aspects) to the original input. (4) cd-GAN works best among all the models by preserving domain-independent features from the source-domain input and combining the domain-specific features from the target-domain conditional input. Here are two examples. (1) In 6-th column of 1-st row, the woman is put on red lipstick. (2) In 6-th column of 5-th row, the hair-style of the generated image is the most similar to the conditional input.

We can get similar observations for women-to-men translations as shown in Figure 3(b), especially for the domain-specific features such as hair style and beard.

From Figure 4 and 5, we find that cd-GAN can well leverage the domain-specific information carried in the conditional inputs and control the generated target-domain images accordingly. DualGAN, DuanGAN-c and GAN-c do not effectively utilize the conditional inputs.

One important characteristic of conditional image-to-image translation model is that it can generate diverse target-domain images for a fixed source-domain image, only if different target-domain images are provided as inputs. To verify such this ability of cd-GAN, we conduct t-

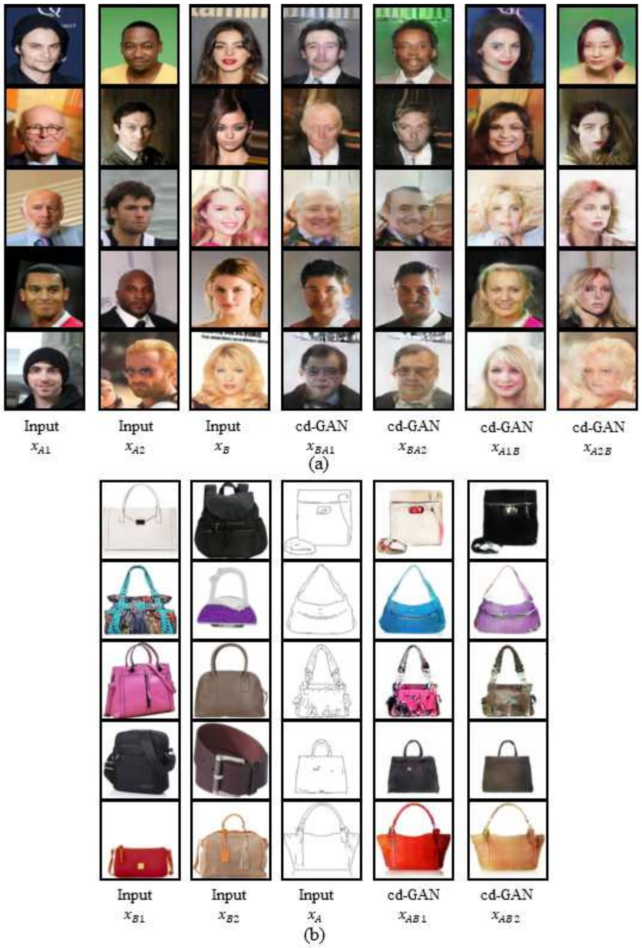

Figure 6. Our cd-GAN model can produce diverse results with different conditional images. (a) Results of women→men translation with two different men's images as conditional inputs. (b) Results of edges→handbags translation with two different handbags as conditional inputs.

wo experiments: (1) for each woman's photo, we work on women-to-men translations with different man's photos as conditional inputs; (2) for each edge of a bag, we work on edges-to-bags translations with different bags as conditional inputs. The results are shown in Figure 6. Figure 6(b) shows that cd-GAN can fulfill edges with the colors and textures provided by the conditional inputs. Besides, cd-GAN also achieves reasonable improvements on most face translations: The domain-independent features like woman's facial outline, orientations and expressions are preserved, while the women specific features like hair-style and the usage the lipstick are replaced with men's. An example is the second row of Figure 6(a), where pointed chins, serious expressions and looking forward are preserved in the generated images. The hairstyles (bald v.s. short hair) and the beard (no beard v.s. short beard) are reflected by the corresponding men's. Similar translations of the other images can also be found.

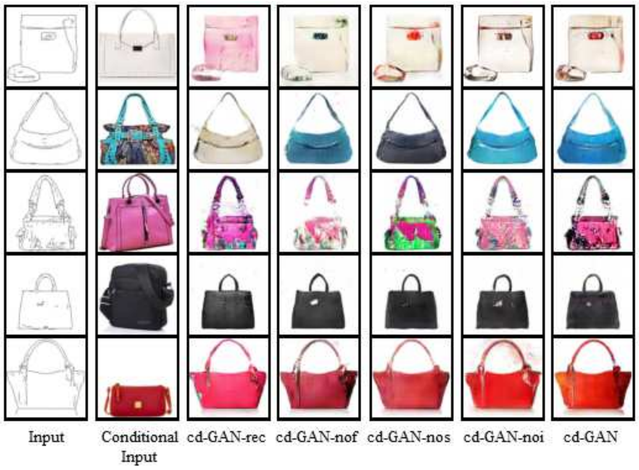

Figure 7. Results produced by different connections and losses of cd-GANs.

Note that there are several failure cases in face translations, such as first column of Figure 6 (a) and last column of Figure 6 (b). Most translated results demonstrate the effectiveness of our model. More examples can be found in our supplementary document.

4.3. Component Study

In this sub section, we study other possible design choices for the model architecture in Figure 2 and losses used in training. We compare cd-GAN with other four models as follows:

- **cd-GAN-rec**. The inputs are reconstructed as

$$\hat{x}_A = g_A(\hat{x}^i_A, \hat{x}^s_A); \hat{x}_B = g_B(\hat{x}^i_B, \hat{x}^s_B) \qquad (12)$$

  instead of Eqn.(7). That is, the connection from x^s_A to g_A in the right box of Figure 2 is replaced by the connection from $\hat{x}^s_A$ to g_A, and the connection from x^s_B to g_B in the right box of Figure 2 is replaced by the connection from $\hat{x}^s_B$ to g_B.

- **cd-GAN-nof**. Both domain-specific and domain-independent feature reconstruction losses, i.e., Eqn.(10) and Eqn.(9), are removed from dual learning losses.

- **cd-GAN-nos**. The domain-specific feature reconstruction loss, i.e., Eqn.(10), is removed from dual learning losses.

- **cd-GAN-noi**. The domain-independent feature reconstruction loss, i.e., Eqn.(10) is removed from dual learning losses.

The comparison experiments are conducted on the edges-to-handbags task. The results are shown in Figure 7. Our cd-GAN outperforms the other four candidate models with better color schemes. Failure of cd-GAN-rec demonstrates the necessity of "skip connections" (i.e., the connections from x_A^s to g_A and from x_B^s to g_B) for image reconstruction. Since the domain-specific feature level and image level reconstruction losses have implicitly put constrains on domain-specific feature to some extent, the results produced by cd-GAN-noi are closest to results of cd-GAN among the four candidate models.

So far, we have shown the translation results of cd-GAN generated from the combination domain-specific features and domain-independent features. One may be interested in what we really learn in the two kinds of features. Here we try to understand them by generating translation results using each kind of features separately:

- We generate an image using the domain-specific features only:

  $$x_{AB}^{A=0} = g_B(x_A^i = \mathbf{0}, x_B^s),$$

  in which we set the domain-independent features to 0.

- We generate an image using the domain-independent features only:

  $$x_{AB}^{B=0} = g_B(x_A^i, x_B^s = \mathbf{0}),$$

  in which we set the domain-specific features to 0.

The results are shown in Figure 8. As we can see, the image $x_{AB}^{A=0}$ has similar style to x_B, which indicates that our cd-GAN can indeed extract domain-specific features. While $x_{AB}^{B=0}$ already loses conditional information of x_B, it still preserves main shape of x_A, which demonstrates that cd-GAN indeed extracts domain-independent features.

4.4. User Study

We have conducted user study to compare domain-specific features similarity between generated images and conditional images. Total 17 subjects (10 males, 7 females, age range $20-35$) from different backgrounds are asked to make comparison of 32 sets of images. We show the subjects source image, conditional image, our result and results from other methods. Then each subject selects generated image most similar to conditional image. The result of user study shows that our model obviously outperforms other methods.

5. Conclusions and Future Work

In this paper, we have studied the problem of conditional image-to-image translation, in which we translate an image from a source domain to a target domain conditioned on another target-domain image as input. We have proposed a new model based on GANs and dual learning. The model can well leverage the conditional inputs to control and diversify the translation results. Experiments on two settings (symmetric translations and asymmetric translations) and three tasks (face-to-face, edges-to-shoes and edges-to-handbags translations) have demonstrated the effectiveness of the proposed model.

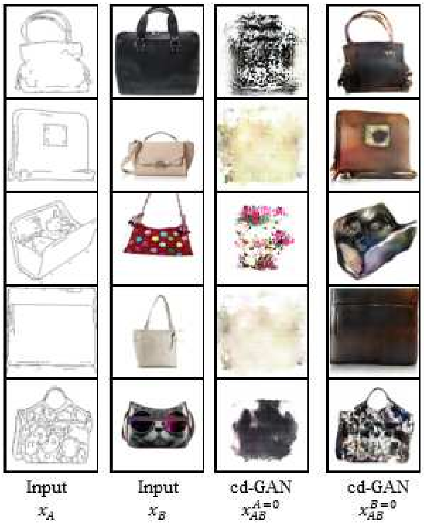

Figure 8. Images generated using only domain-independent features or domain-specific features.

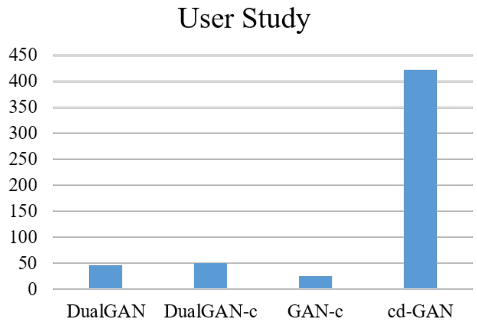

Figure 9. The result of user study.

There are multiple aspects to explore for conditional image translation. First, we will apply the proposed model to more image translation tasks. Second, it is interesting to design better models for this translation problem. Third, the problem of conditional translations may be extend to other applications, such as conditional video translations and conditional text translations.

6. Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant No.2016YFC0801001, NSFC under Grant 61571413, 61632001, 61390514, and Intel ICRI MNC.

References

[1] E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.

[2] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[3] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

[4] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[5] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, 2012.

[6] 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.

[7] 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.

[8] 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.

[9] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

[10] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[12] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[13] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[14] Y. Lu, Y.-W. Tai, and C.-K. Tang. Conditional cyclegan for attribute guided face image generation. *arXiv preprint arXiv:1705.09966*, 2017.

[15] P. Luo, G. Wang, L. Lin, and X. Wang. Deep dual learning for semantic image segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[16] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013.

[17] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. ICML*, pages 807–814, 2010.

[18] Y. Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

[19] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[20] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

[21] Y. Xia, T. Qin, W. Chen, J. Bian, N. Yu, and T.-Y. Liu. Dual supervised learning. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3789–3798, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[22] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[23] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 192–199, 2014.

[24] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[25] 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.