# One-Shot Image-to-Image Translation via Part-Global Learning with a Multi-adversarial Framework

Ziqiang Zheng, Zhibin Yu, *Member, IEEE,* Haiyong Zheng, *Member, IEEE,* Yang Yang, *Senior Member, IEEE,* and Heng Tao Shen, *Senior Member, IEEE*

***Abstract*—It is well known that humans can learn and recognize objects effectively from several limited image samples. However, learning from just a few images is still a tremendous challenge for existing main-stream deep neural networks. Inspired by analogical reasoning in the human mind, a feasible strategy is to "translate" the abundant images of a rich source domain to enrich the relevant yet different target domain with insufficient image data. To achieve this goal, we propose a novel, effective multi-adversarial framework (MA) based on part-global learning, which accomplishes the one-shot cross-domain image-to-image translation. In specific, we first devise a part-global adversarial training scheme to provide an efficient way for feature extraction and prevent discriminators from being overfitted. Then, a multi-adversarial mechanism is employed to enhance the image-to-image translation ability to unearth the high-level semantic representation. Moreover, a balanced adversarial loss function is presented, which aims to balance the training data and stabilize the training process. Extensive experiments demonstrate that the proposed approach can obtain impressive results on various datasets between two extremely imbalanced image domains and outperform state-of-the-art methods on one-shot image-to-image translation. Our code will be released with this paper at https://github.com/zhengziqiang/OST.**

***Index Terms*—One-shot, generative adversarial networks, image-to-image translation, unpaired cross-domain translation.**

## I. Introduction

BENEFITED from the great success of deep learning-based approaches, researchers have made much progress on computer vision fields such as image classification [1], [2], [3], [4], [5], [6], [7], image retrieval [8], [9], [10], [11], [12], [13], [14], [15], [16] and image hashing [17], [18], [19], [20], [21], [22], [23]. Generally, these methods may achieve reasonable results based on sufficient data for training the deep neural networks [3], [4]. However, collecting and labeling those data are time-expensive and tedious. In certain real-world scenarios, it may be impossible to gather abundant data from the target domain $Y$ because of the scarcity of the image samples (In the worst case, it is possible to have only one image from $Y$). Nonetheless, we probably have redundant data from another source domain $X$, whose image samples are correlated to the ones in the target domain $Y$ (such as photo and sketch images as shown in Fig. 1). It would be a feasible solution if we generate images of the domain $Y$ corresponding to analogous images of the domain $X$ based on the diversity while keeping the semantic matching. So it is necessary to obtain efficient intermediate representation from limited samples.

Previous one-shot work mainly focuses on one-shot image recognition [24], [25], [26], [27], [28], [29]. They try to find a meta-learning framework, which could easily adapt to a new task with slight fine-tuning on one sample. However, these methods are not robust, the trained recognition model usually gives a totally different answer if the pose information of objects has changed. In order to improve the robustness of the one-shot visual recognition task, some traditional data augmentation methods are adopted, including the random flipping, rotation, and the cropping operations [25], [30], these methods do not improve the content diversity of the one-shot sample, they only achieve a few improvements [25], [30]. In this paper, we mainly concern about one-shot unpaired image-to-image translation. Our purpose is to find a mapping function $\mathcal{F}$ to translate images from the source image domain $X$ to the target image domain $Y$ with only one image sample as shown in Fig. 1. Using image translation, we can enrich training samples of the target domain through translating images from a relevant source domain even if limited target domain samples are given. After the data augmentation through the one-shot image translation, the one-shot image translation methods can be easily combined with previous one-shot visual tasks and there is a great potential to improve the recognition accuracy by combining the generative models and the one-shot recognition models. So the one-shot image translation is worthy to be explored.

Concerning the image-to-image translation field, Gayts *et al.* [31] first proposed a "Neural Style" algorithm, which combines the content of one image with the style of another image using convolutional neural networks. Johnson *et al.* [32] adopted the perceptual distance to measure the content and style similarity between different images. However, the translation result is confined to image painting style translation without high-level semantic matching. Advanced by the powerful ability of modeling visual content of generative adversarial networks (GANs), several recent research endeavors

Z. Zheng, Y. Yang, and H.T. Shen are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. (e-mail: zhengziqiang1@gmail.com; dlyyang@gmail.com; shenhengtao@hotmail.com).

Y. Yang is also with Institute of Electronic and Information Engineering of UESTC in Guangdong.

Z. Yu and H. Zheng are with the School of Electronic information engineering, Ocean University of China. (e-mail: yuzhibin@ouc.edu.cn; zhenghaiyong@ouc.edu.cn).

Corresponding author: Yang Yang.

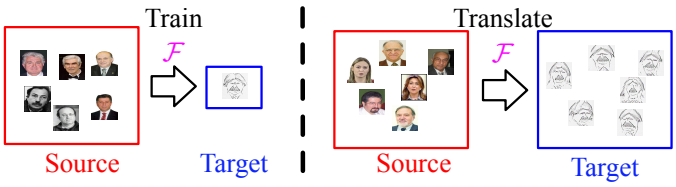



Fig. 1. One-shot cross-domain image-to-image translation. Note that we have only one target sample for training.

have been devoted to applying adversarial training [33] to enhance the robustness and generality of traditional image-to-image translation [34], [35], [36], [37], [38], [39], [40], [41]. Guo *et al.* [42] adopted the auto-embedding to achieve high-resolution image synthesis. Qiu *et al.* proposed a novel adversarial semantic segmentation architecture to handle pixel-level image understanding. These methods can obtain acceptable performance by using sufficient training data from both the target and the source domains. As aforementioned, we usually encounter the situation that the target domain does not have enough training samples. In certain cases, we only have one sample, which even has no counterpart in the source domain.

The first attempt about one-shot unpaired cross-domain image-to-image translation [43] focuses on a one-to-many image-to-image translation scheme, which transforms the only one target sample into the source domain. Such an assimilation process may easily cause the loss of the specific knowledge attached to the target sample. Obviously, this method can not be applied for the data augmentation of the target domain. In contrast, we target at the many-to-one image-to-image translation, i.e., converting the diverse source samples to the target domain. We argue this is more challenging due to the extremely limited exploration of the target domain. To overcome the above obstacle, one feasible solution is to exploit the generative power of GANs to enable many-to-one translation. Nevertheless, direct applying GANs may suffer from two major challenges: 1) the imbalance of insufficient target data and abundant source data leading to the overfitting during the learning process of the discriminator in the target domain; and 2) the lack of discriminative ability for extracting the high-level semantic representation, thereby failing to transfer the semantic information from the source domain to the target image domain.

Intuitively, learning from one sample in the human mind usually relies on part-to-part analogical reasoning to obtain fine-grained information. Inspired by such a process, in this work, we devise a part-based discriminator procedure, which is capable of distinguishing the local part randomly cropped from translated images and real images using the limited information from the target domain. The benefits of designing the part-based discriminator are two-fold: 1) it helps to capture local characteristics of the target samples in a more accurate manner, and 2) it can assist the discriminator in alleviating the overfitting problem by using random partial information rather than the entire image. Besides, to balance the learning for the target data and the source data, we devise a balanced adversarial loss, which utilizes a controlling hyper-parameter to reduce the convergence speed of the objective function. It is worth noting that if we do not use this balanced adversarial loss, the model tends to suffer from the trivial solution and cause the overfitting problem, i.e., generating extremely similar samples to the only one target sample no matter what the input of the source domain is. Furthermore, following [40], we divide an original discriminator into a bunch of weak learners via multiple threads, which not only helps to significantly improve the efficiency by reducing the number of the training parameters but also digs in more fine-grained semantic details of the only one target sample.

The contribution of this paper is summarized as follows:

- We propose a novel and effective one-shot image-to-image translation framework to translate abundant images from a source domain to another target image domain containing only one image. To the best of our knowledge, our work is one of the first attempts to achieve the one-shot unpaired cross-domain image-to-image translation in the many-to-one setting.
- We propose to utilize a multi-adversarial mechanism via part-global learning to enhance the ability of the discriminator in characterizing the fine-grained semantics as well as significantly improve the efficiency of the training process.
- We introduce a balanced adversarial loss function to alleviate the influence of the data imbalance between the target domain and the source domain.

The rest of this paper is organized as below. Section II briefly introduces the related work and Section III elaborates the proposed approach. Section IV presents the extensive experimental results on various datasets, followed by the conclusion in Section V.

## II. Related Work

### A. Image-to-image translation

Due to the success of conditional GAN [44], many popular image-to-image translation methods were developed such as Pix2pix [39] and Pix2pixHD [45]. They could achieve high-resolution and precise image synthesis [42] by training with paired images. However, paired training data are not always available. To overcome this shortage, many unpaired image domain translation were proposed including CycleGAN [34], DualGAN [35], DiscoGAN [46], UNIT [37], MUNIT [36], DRIT [38] and so on [47]. These methods could translate images from one domain to another domain based on unpaired images. CycleGAN [34] adopted a cycle-consistency loss to constrain the reconstruction of target images. MUNIT [36] adopted an unsupervised multimodal structure to translate styles as well as contents to reconstruct the target images. The concurrent DRIT [38] aimed to generate images with diverse outputs, which proposed a disentangled representation framework. GANimorph [48] combined shape deformation and dilated convolutions to perform cross-species translation. Besides, Twin-GAN [49] used a progressively growing skip connected encoder-generator structure for human-anime character translation. Nevertheless, most of these works mainly performed experiments with redundant images from both

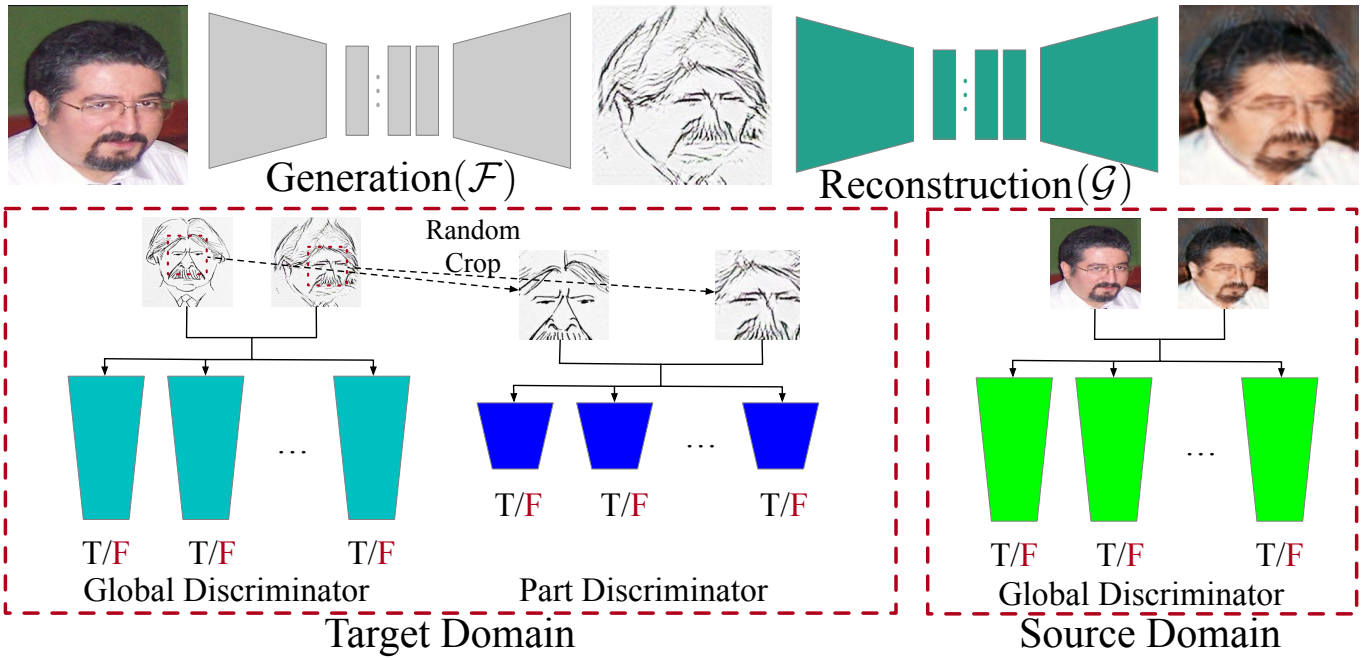



Fig. 2. The framework of our method. We take advantage of part-global learning together with a multi-adversarial discriminator architecture (MA). $\mathcal{F}$ aims to translate images from source domain to target domain while $\mathcal{G}$ tries to reconstruct the outputs of $\mathcal{F}$ in source domain.

source and target domains, which could perform unsatisfactorily when limited images are given.

### B. One-shot image translation

One-shot learning, which was first discussed by Fei-Fei Li and Erik Miller [28], [29], aims to learn information about object categories from one, or only a few, training samples. Most of the published one-shot learning approaches focus on how to recognize objects from a few samples (one sample) [50], [24], [51], [25], [30]. Different from above one-shot object recognition methods, one-shot image translation (OST) aimed to translate images between two domains in which one domain only includes one or a few images. This concept was first discussed by Benaim *et al.* [43], who aimed to generate an analogous of $y$ in $X$, with a single image $y$ from domain $Y$ and a set of images from domain $X$. To find a mapping function, they shared some specific layers of one variational autoencoder [52] to add a strong constraint between domain translation. Unlike their task, we aim to perform a more challenging task, which discovers a semantic mapping function to translate a set of images from $X$ to $Y$, namely, we have reverse translation direction with OST methods [43]. In our case, we aim to use the semantic link between domains and unearth perceptual similarity between $X$ and $Y$ with abundant images of $X$, and one sample image of $Y$ is given. Currently, SinGAN [53] is developed to synthesize the plausible multi-scale patch images with only one natural image. It aims to manipulate the repetitive pattern features from the one-shot target sample and the source content information is not introduced during the synthesis stage. Differently, we target to achieve the data augmentation of the target domain at the one-shot setting through the image translation.

### C. Multi-adversarial training

Recently many methods have utilized multi-adversarial training mechanisms to enhance generation performance, which ensemble different discriminators functionally. GMAN [54] first adopted multiple discriminators for high-quality image generation with fast and stable convergence. Multi-discriminator CycleGAN [55], which is an extension of CycleGAN, was proposed to enhance the speech domain adaption with a multiple discriminators architecture. MD-GAN [56] was proposed to use a GAN with multiple discriminators on the distributed datasets. Most of the studies have used multiple discriminators to give the generator with better guidance. Pix2pixHD [45] and MUNIT [36] adopted multi-scale discriminator structure for high-resolution paired and multimodal unpaired image-to-image translation respectively. And Yang et al [57] also apply multi-task learning to explore the internal correlation. Recently GAN-MBD [40] proposed a multi-branch discriminator to reduce the parameter of discriminators and enhance the translation between species. Based on the multi-adversarial training, the image generation and translation quality have made comprehensive progress. For our purpose, with limited images given, we aim to use the multi-adversarial training mechanism to improve the image-to-image translation process and increase the possibility to establish a high-level semantic link between different domains.

## III. The Proposed Approach

In this section, we elaborate the proposed approach for many-to-one image-to-image translation.

### A. Part-Global discriminators

As illustrated in Fig. 3, suppose we want to translate images from the source "cat" domain to the target "dog" domain with redundant "cat" samples and only one "dog" sample, the intuitive principle of analogical reasoning is to 1) preserve global layout/pose of the original image, as well as 2) perform semantic matching of detailed parts, such as eyes, ears, and nose. Iizuka *et al.* [58] proposed a global-local adversarial architecture to effectively combine global and local information to boost image inpainting performance. In particular, a local context discriminator was proposed to ensure local consistency, which makes sure the input of this local discriminator is a small area centered at the completed region. Inspired by the powerful modeling ability of local and global information, we devise a part-global adversarial architecture to increase the variety of the source domain and improve the one-shot image-to-image translation process. Specifically, our part discriminator is fed with a random part cropped from generated images and real images as shown in Fig. 2. It is worth noting that our part discriminator is only designed for generator $\mathcal{F}$ in the target domain. Furthermore, we only feed a small part that randomly cropped from the entire real/fake image to enhance the robustness of models and improve the ability to discover the target representation. Through this method, more fine-grained part samples could be reachable by random cropping, thus our model could capture more detailed information from local context parts. And it can also alleviate the over-fitting problem at the one-shot setting. To ensure the global consistency and semantic matching between generated images and the only one target sample image, we combine a common global discriminator to cope with the entire images. Note that we consider part discriminator and global discriminator as an equal contribution to the learning process. The loss function can be described as:

$$\mathcal{L}(\mathcal{F}, D) = \begin{cases} \mathcal{L}_p(\mathcal{F}, D_p) & \text{for } D_p, \\ \mathcal{L}_g(\mathcal{F}, D_g) & \text{for } D_g, \end{cases} \tag{1}$$

where

$$\begin{aligned} \mathcal{L}_g(\mathcal{F}, D_g) =& \mathbb{E}_{y \sim p_{\text{data}}(Y)}[\log D_g(y)] + \\ & \mathbb{E}_{x \sim p_{\text{data}}(X)}[\log(1 - D_g(\mathcal{F}(x)))], \end{aligned} \tag{2}$$

and

$$\begin{aligned} \mathcal{L}_p(\mathcal{F}, D_p) =& \mathbb{E}_{\hat{y} \sim p_{\text{data}}(Y)}[\log D_p(\hat{y})] + \\ & \mathbb{E}_{\hat{x} \sim p_{\text{data}}(X)}[\log(1 - D_g(\mathcal{F}(\hat{x})))]. \end{aligned} \tag{3}$$

Here $D_p$ and $D_g$ denote part discriminator and global discriminator respectively, $\mathcal{F}(\hat{x})$ and $\hat{y}$ denote the random part region cropped from generated images and real images respectively. For both the global and part discriminators, we adopt the PatchGAN architecture. We comprehensively consider that different tasks with different kinds of images (e.g., facial images, natural scenes) might influence the performance of our patch-wise adversarial learning, although we try to make it work on more cases as a general-purpose solution. However, as the fundamental challenge of one-shot learning, lacking of sufficient knowledge makes almost all the current state-of-the-art image-to-image translation methods fail to generalize to all possible cases. Our method benefits from the part-global patch-wise learning to discover the semantic mapping on the many-to-one image-to-image translation tasks.

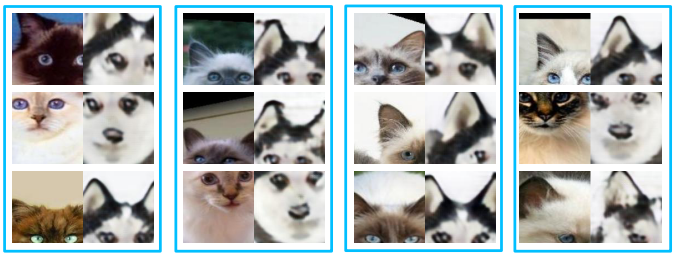


Fig. 3. Illustrative exemplars of translating "cat" to "dog" using proposed approach, which mimics analogical reasoning in the human mind while preserving global layout/pose information and achieving local semantic matching.

### B. Characterizing fine-grained semantics

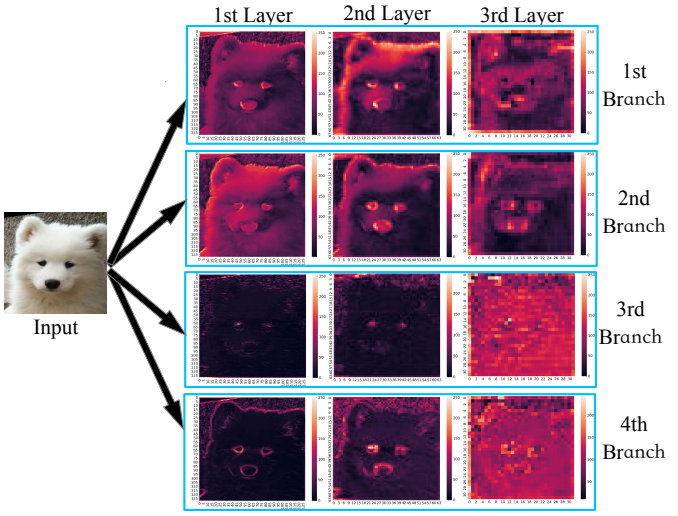



Fig. 4. The visualization of multi-adversarial discriminators for the only one target domain sample. Different threads capture different semantic details.

To characterize the detailed semantics in the image, we propose to utilize the "divide-and-conquer" strategy, i.e., designing different threads of discriminators to decide whether the current image and part region are real or synthesized. The formulation can be described as:

$$\begin{aligned} \mathcal{L}(\mathcal{F}, D) =& \frac{1}{N} \sum_{i}^{N} \{\mathbb{E}_{y \sim P_{\text{data}}(Y)}[\log D_i(y)] \\ & + \mathbb{E}_{x \sim P_{\text{data}}(X)}[\log(1 - D_i(\mathcal{F}(x))]\}, \end{aligned} \tag{4}$$

where $N$ denotes the thread number of a common discriminator. Specifically, we break the whole discriminator into multiple smaller threads by channels, thus the complexity of architecture, i.e., the parameters of discriminators, can be reduced. We feed the average adversarial loss of discriminators to update generators, and each thread of discriminators is optimized independently. As described in [40], each thread of discriminators could learn a semantic sub-task automatically. With this implicit semantic division, our model could have a stronger ability to unearth the intrinsic link between the

source and target domains. As illustrated in Fig. 4, to make an explicit elaboration, we train discriminators of 4 threads with abundant samples from the source domain and the only one target sample and visualize the feature map outputs of the only one target sample. Each thread of the model can take charge of a different semantic representation of the only one target sample. The first thread focuses on eyes and nose while the second thread pays attention to fur information. The third thread captures small detailed information, and the fourth thread observes the edge information. To be noted, we do not provide explicit additional constrains between the threads of the discriminator, and the implicit semantic division is learned during the whole training procedure based on the training samples. Even there is no explicit loss between the multiple branches of the target discriminator, the average loss of the multiple branches is fed into the generator, and each branch constitutes the adversarial training with the generator, so there is an implicit connection between the different threads.

### C. Balance training between source and target

In consideration of the possible extreme imbalance between the target and the source domains for *one vs. many* case, the discriminators of the target domain could be easily over-fitted if we keep the same training speed between discriminators of the target and source domains. To alleviate this problem, we develop a balanced adversarial loss to slow down the convergence to the one-shot image. Here we design a strategy by using a hyper-parameters $\alpha$ to control the convergence speed of discriminators for two domains, which can boost the convergence of the source and target discriminators simultaneously. For the mapping function $\mathcal{F}: X \rightarrow Y$ and $\mathcal{G}: Y \rightarrow X$, the balanced adversarial loss is defined as:

$$\mathcal{L} = \alpha \mathcal{L}_{(\mathcal{F}, D_Y)} + \mathcal{L}_{(\mathcal{G}, D_X)}, \tag{5}$$

where $D_X$ and $D_Y$ denote discriminators for source domain and target domain respectively. We have explored the effectiveness of this hyper-parameter in our framework and implemented various experiments using different values of $\alpha$ in Section IV-E.

## IV. Experiments

### A. Datasets

We evaluate our approach by comparing with the state-of-the-art on six different datasets:

**Caricature** [59] includes 200 paired caricature images, which deforms the facial feature of real images.

**IIIT-CFW** [60] contains 1000 real images and 8928 annotated cartoon faces of famous characters of the world with a varying profession of 100 public figures.

**CelebA+Portrait** [38] is a combined dataset derived from CelebA [61] and Wikiart[1]. In specific, 6453 images are selected from CelebA as the source domain, and 1814 images are selected from Wikiart as the target domain.

**Cat2dog** is a cropped image dataset including 871 cat images and 1364 dog images in total. We inherit this dataset from DRIT [38], and we follow the same data split for training and testing.

**Day2night** [62] contains 100 paired day-night images in which 1+90 (*one vs. many*) images are used for training.

**PHOTO-SKETCH** [63], [62] is a photo-to-sketch translation dataset, which contains paired facial photos and sketch images.

[1]https://www.wikiart.org/

### B. Implementation details

We mainly inherit the architecture from CycleGAN [34]. We extend the layers of discriminators with part-global discriminators to capture the high-level semantic representation and adopt the multi-adversarial training mechanism in our model. The part discriminators own fewer layers than global discriminators. Inspired by the PatchGAN of Pix2pix proposed in [39], we derive the global and part discriminators. But we use different architectures of a 5-layer global discriminator and a 4-layer part discriminator, yielding 32*32 outputs and 8*8 outputs, respectively. For each layer of the discriminators, a convolutional layer with kernel size 4 and stride 2 is included, and we add one Leaky ReLU with slope 0.02 after the Conv layer. For the generator architecture, we adopt three Conv-InstanceNorm-ReLU blocks to achieve a downsample of the input images, the kernel size is set to 3 and the stride is 2. For the bottleneck, 9 residual blocks are applied to stack the content information. 3 Deconv-InstanceNorm-ReLU blocks are performed to generate the same size image outputs. Finally, a Tanh activity function is applied to obtain the normed outputs. To improve the generality and robustness of models, we use some common data augmentation approaches including random flip, slightly rotation, and center crop. The hyper-parameter $\alpha$ mentioned in Eq. 5 is set as $0.1$ in all our experiments. We use Adam [64] to optimize our model and set the learning rate at $0.0002$.

### C. Evaluation metrics

To evaluate the effectiveness of different methods, we measure the translated quality by using the following criteria:

**Fréchet Inception Distance** (FID) [65] computes the similarity between the generated sample distribution and real data distribution. This method is a consistent and robust approach for evaluating the generated images [66], [67], and it can be calculated by:

$$\text{FID} = ||\mu_x - \mu_g||_2^2 + Tr\left(\sum_x + \sum_g - 2(\sum_x \sum_g)^{\frac{1}{2}}\right), \tag{6}$$

where $(\mu_x, \sum_x)$ and $(\mu_g, \sum_g)$ are the mean and covariance of the sample embeddings from the data distribution and model distribution, respectively. Lower FID index means that the smaller distribution difference between the generated and the target images, and which represents higher generated image quality. In our one-shot unpaired image-to-image translation task, we can evaluate the image generation quality by computing these metrics.

**Learned Perceptual Image Patch Similarity** (LPIPS) [68] computes the perceptual similarity between two images. A lower LPIPS means that the two images have more perceptual similarity. Considering two image domains, we can compute

the LPIPS distance to evaluate the perceptual similarity.

**Structural Similarity** (SSIM) [69] is a traditional metric to measure the similarity between two images. Higher SSIM shows more structural similarity between generated images and real images.

### *D. Comparison with the state-of-the-arts*

We compare our method to state-of-the-art image-to-image translation methods: CycleGAN [34], MUNIT [36] and DRIT [38]. The comparison is performed under two settings: *one vs. many* and *many vs. many*. For *one vs. many* case, we use only one image from the target domain and many images from the source domain. For *many vs. many* case, we use many images from both the two domains. We also compare our method with the OST [43] method and fast-neural-style [32] method for *one vs. many* case.

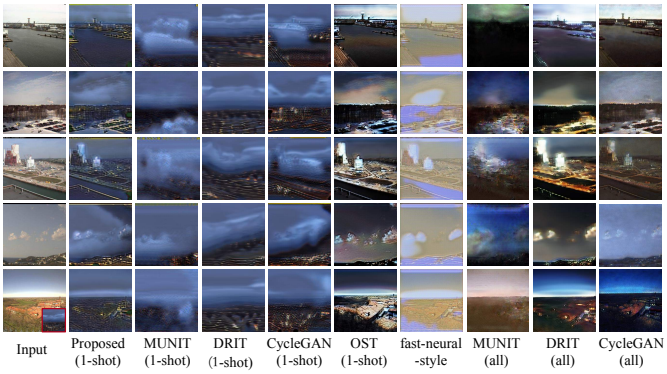



Fig. 5. The day→night translation results on the day2night dataset using different methods. The smaller image framed by the red box in the lower-left corner input shows the only one training sample from the target domain.

*1) Results on scene change:* We first conduct a scene change task on **day2night** [62] dataset. Fig. 5 shows the translated results. It can be seen, our method generates satisfactory nighttime images from daytime inputs while using only one night time image as the training sample for *one vs. many* case and our method can preserve the perceptual contents of input images. In contrast, DRIT generates images with several dirty color blocks and fails to generate reasonable objects. MUNIT is nearly overfitted with the only one training sample and the generated results are similar to the one-shot sample no matter what the input looks like. Both DRIT and MUNIT fails to achieve plausible translation for *one vs. many* case. The fast-neural-style method merely achieves the color and textile translation, yielding unnatural synthesized images. For quantitative comparison, we compute the quantitative results (FID, LPIPS, and SSIM) between outputs and paired ground truth images provided in this dataset, and results are listed in Tab. I. Our method has the lowest FID and highest SSIM value among all the methods for *one vs. many* case, which indicates the better image translation performance and stronger ability to preserve the structure information of input samples.

*2) Results on photo-to-caricature:* In this part, we evaluate our method on a more challenging task, which aims to achieve the photo-to-caricature translation. Gats *et al.* [31] and Jonhson *et al.* [32] performed artist style transformation using one style image and a set of input images. We conduct experiments on four photo-to-caricature datasets, i.e., Caricature [59], PHOTO-SKETCH [63], [70], IIIT-CFW [60] and CelebA+Portrait [38]. This task requires not only satire exaggeration of photos but also artist style transfer. We compare our method to others under two settings: *one vs. many* and *many vs. many*.

TABLE I
QUANTITATIVE COMPARISON OF DIFFERENT METHODS FOR DAY→NIGHT TRANSLATION TASK ON DAY2NIGHT DATASET.

| Method | FID | LPIPS | SSIM |
|---|---|---|---|
| Proposed (1-shot) | **223.4** | 0.6887 | **0.6959** |
| DRIT (1-shot) | 290.4 | 0.6913 | 0.6917 |
| MUNIT (1-shot) | 261.8 | **0.6675** | 0.6636 |
| CycleGAN (1-shot) | 248.9 | 0.6783 | 0.6823 |
| OST (1-shot) | 355.3 | 0.7063 | 0.5748 |
| fast-neural-style | 301.0 | 0.7430 | 0.4672 |
| DRIT (all) | 227.0 | 0.6646 | 0.5555 |
| MUNIT (all) | 225.2 | 0.6654 | 0.5811 |
| CycleGAN (all) | 327.0 | 0.6885 | 0.7174 |

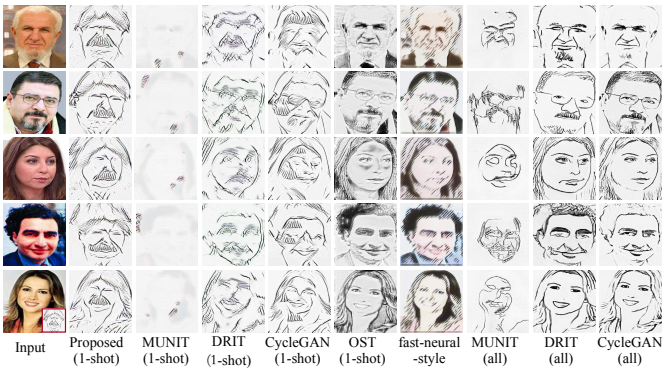



Fig. 6. The photo→caricature translation results on Caricature dataset using different methods. The smaller image framed by red box in lower left corner input shows the only one training sample from target domain.

TABLE II
QUANTITATIVE COMPARISON OF DIFFERENT METHODS FOR PHOTO→CARICATURE TRANSLATION TASK ON CARICATURE DATASET.

| Method | FID | LPIPS | SSIM |
|---|---|---|---|
| Proposed (1-shot) | **202.6208** | **0.5523** | **0.9629** |
| DRIT (1-shot) | 247.7 | 0.5823 | 0.9577 |
| MUNIT (1-shot) | 312.0 | 0.7211 | 0.9616 |
| CycleGAN (1-shot) | 272.3 | 0.5813 | 0.9604 |
| OST (1-shot) | 304.7 | 0.6762 | 0.9076 |
| fast-neural-style | 278.7 | 0.6507 | 0.8554 |
| DRIT (all) | 97.82 | 0.5506 | 0.9491 |
| MUNIT (all) | 144.8 | 0.6019 | 0.9552 |
| CycleGAN (all) | 133.4 | 0.5445 | 0.9547 |

Fig. 6 reports the translation results using different methods on the Caricature dataset. The *one vs. many* task used only one randomly-selected caricature image as the target sample and 160 photos, while the *many vs. many* tasks used 160 photo images and 160 caricature images. The rest 40 pairs were used

for testing. For the *one vs. many* case, our method not only captures the caricature style but also preserves the pose, the layout, and identity information of inputs. As can be observed from the second column of Fig. 6, our method generates exaggerated mustaches on the appropriate part. For *one vs. many* case, CycleGAN only generates a mustaches artifact on the same part region for all the generated samples. DRIT generates images with blur boundary and some artifacts, while MUNIT fails to synthesize satisfactory results. OST and fast-neural-style methods only obtain colored outputs according to the only one target sample without caricature translation. Since our one-shot setting of image-to-image translation only provides one image of a domain for training, it is extremely challenging for models to learn domain knowledge and common sense (women don't have mustaches), so all the generated images by our method have the mustaches at the reasonable position. With only one caricature image given, our model has captured the mustaches representation (all the humans have mustaches) and the relationship between different parts (the mustaches should be under the nose, above the mouth). This also happens to our human, considering that a baby has only seen people with mustaches (described by the one-shot target sample), it will be normal for the baby to think that people should have mustaches. Thus, it makes sense of our outputs in Fig. 6.

Tab. II presents the FID, LPIPS, and SSIM values between generated images and ground truth images for all the evaluated methods. Our method achieves the lowest LPIPS distance and highest SSIM score compared to the other methods. The lowest LPIPS represents the smaller perceptual distance between the generated images and real images. Compared with methods trained using all the training samples at *many vs. many* case, our method can even obtain a higher SSIM score.

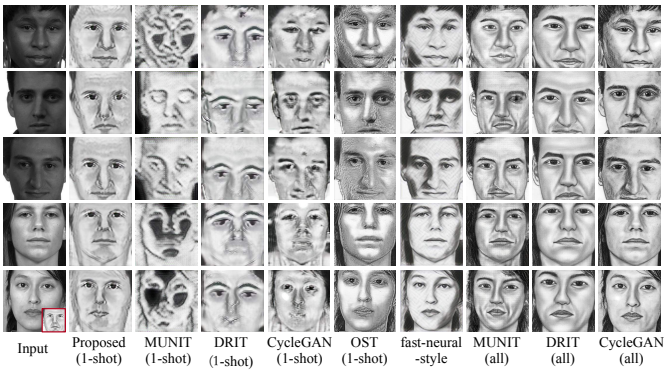



Fig. 7. The photo→sketch translation results on PHOTO-SKETCH dataset using different methods. The smaller image framed by the red box in the lower-left corner input shows the only one training sample from the target domain.

Fig. 7 illustrates the translated results of different approaches on PHOTO-SKETCH [63], [70] dataset, which has consistent sketch style. Tab. III shows quantitative performance of different methods. We used 995 photos and one randomly selected sketch image for the source domain and the target domain, respectively. The rest 199 photo-sketch image pairs are used for testing. As seen, all the methods can get plausible results when using all 995 training paired images. Nonetheless, when only one target sample is fed, most of them achieve poor

TABLE III
QUANTITATIVE COMPARISON OF DIFFERENT METHODS FOR PHOTO→SKETCH TRANSLATION TASK ON PHOTO-SKETCH DATASET.

| Method | FID | LPIPS | SSIM |
|---|---|---|---|
| Proposed (1-shot) | **94.28** | 0.3995 | **0.9172** |
| DRIT (1-shot) | 115.7 | 0.5400 | 0.9037 |
| MUNIT (1-shot) | 370.3 | 0.5452 | 0.8127 |
| CycleGAN (1-shot) | 192.4 | 0.5941 | 0.9086 |
| OST (1-shot) | 141.8 | 0.4846 | 0.9140 |
| fast-neural-style | 143.1 | **0.3603** | 0.9076 |
| DRIT (all) | 25.93 | 0.2418 | 0.9554 |
| MUNIT (all) | 41.19 | 0.2782 | 0.9392 |
| CycleGAN (all) | 34.52 | 0.2478 | 0.9471 |

performance. In contrast, our method, which gains the lowest LPIPS, the lowest FID, and the highest SSIM performance, can well preserve the pose/layout information of the source samples and generate vivid sketch similar to the only one target style.

Further, we conducted experiments on IIIT-CFW [60] dataset. We randomly selected one image from the caricature domain as a training target and 800 photo images as training sources. For *many vs. many* case, we used all the caricature images as the target. The rest 200 photo images are used for testing. Since IIIT-CFW does not provide real-cartoon image pairs, we only report the FID performance between the synthesized images and the real images as illustrated in Tab. IV. Our approach method achieves the best results among all the evaluated methods in *one vs. many* case. As illustrated in Fig. 8, compared to other methods under *one vs. many* that merely perform the textile transformation, our method can better characterize the semantic aspects (e.g., eyes and eyebrows), preserve the layout/pose information from source inputs, as well as inherit the style from the only one target sample.

We also perform portrait translation using CelebA+Portrait [38] dataset. We follow the training/testing set as in [38]. Since the source photos and the portrait images are not paired, we only compute the FID performance. We report the quantitative comparison results in Tab. IV, from which we can observe that our proposed approach outperforms other competitors. Furthermore, as illustrated in Fig. 9, in *one vs. many* case, the compared methods either encounter overfitting problem with almost the same outputs for all the source inputs (i.e., DRIT) or achieve unacceptable translation results (i.e., MUNIT). In contrast, our method can effectively preserve semantic details of source inputs (e.g., accessories, glasses), as well as the global layout/pose knowledge.

*3) Results on cat↔dog:* We also evaluate our model on the cat2dog dataset, which is a more challenging task to perform cross-species image-to-image translation. Fig. 10 shows the cat→dog translation results using different methods (only one dog image). Our method can preserve the layout/pose information and achieve the feature matching in high-level space. The OST method fails to achieve cross-species semantic translation. The MUNIT (1-shot) method achieves unsatisfac-

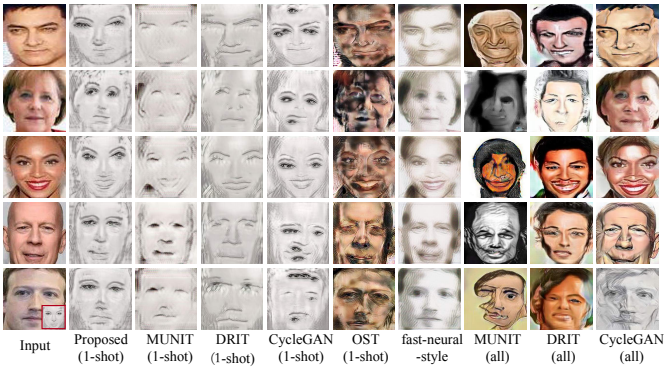



Fig. 8. The photo→caricature translation results on IIIT-CFW dataset using different methods. The smaller image framed by red box in lower left corner input shows the only one training sample from target domain.

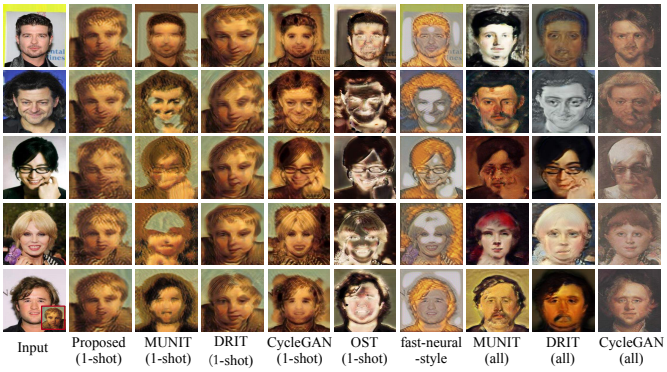



Fig. 9. The photo→portrait translation results on CelebA+Portrait using different methods. The smaller image framed by red box in lower left corner input shows the only one training sample from target domain.

TABLE IV
FID SCORES OF DIFFERENT METHODS ON PHOTO→CARICATURE TRANSLATION TASK ON IIIT-CFW AND CELEBA+PORTRAIT DATASETS. SMALLER FID SCORES SHOW BETTER TRANSLATION QUALITY BETWEEN OUTPUTS AND REAL IMAGES FROM TARGET DOMAIN.

| Method | IIIT-CFW | CelebA+Portrait |
|---|---|---|
| Proposed (1-shot) | **144.7** | **144.5** |
| DRIT (1-shot) | 178.5 | 183.6 |
| MUNIT (1-shot) | 263.3 | 248.2 |
| CycleGAN (1-shot) | 178.4 | 156.6 |
| OST (1-shot) | 245.2 | 179.5 |
| fast-neural-style | 166.4 | 278.7 |
| DRIT (all) | 121.2 | 139.2 |
| MUNIT (all) | 139.6 | 130.6 |
| CycleGAN (all) | 115.6 | 131.5 |

tory results and generates imprecise semantic features, such as nose and eyes. Compared to DRIT (1-shot), our method can obtain better results and more detailed information. Contrary to Fig. 10, Fig. 11 shows the dog→cat translation results (only one cat image). The MUNIT (1-shot) method is unable to achieve precise translation while DRIT (1-shot) method may lead to an overfitting problem, thereby generating similar images with the one training image. Compared to the methods trained using all images, our method cannot generate various backgrounds. Tab. V shows the FID results using all methods.

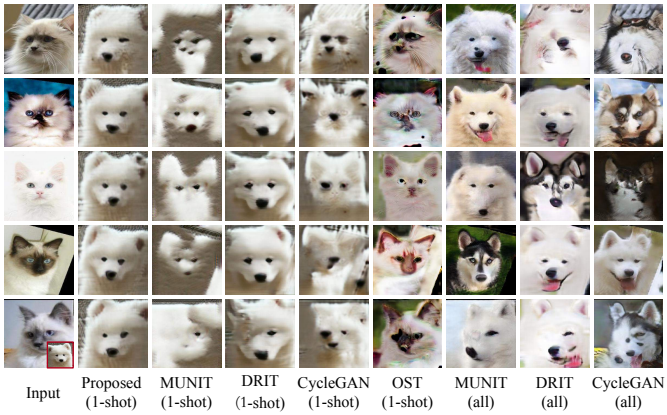



Fig. 10. The cat→dog translation results on cat2dog dataset using different methods. The smaller image framed by red box in lower left corner input shows the only one dog image from target domain.


Fig. 11. The dog→cat translation results on cat2dog dataset using different methods. The smaller image framed by red box in lower left corner input shows the only one cat image from target domain.

TABLE V
FID SCORES OF DIFFERENT METHODS FOR CAT↔DOG TRANSLATION TASKS ON CAT2DOG DATASET.

| Method | dog to cat | cat to dog |
|---|---|---|
| Proposed (1-shot) | **124.8** | **66.64** |
| DRIT (1-shot) | 127.7 | 123.7 |
| MUNIT (1-shot) | 212.2 | 222.1 |
| CycleGAN (1-shot) | 155.1 | 261.5 |
| OST (1-shot) | 277.1 | 324.0 |
| DRIT (all) | 58.44 | 88.63 |
| MUNIT (all) | 46.29 | 47.41 |
| CycleGAN (all) | 46.97 | 64.40 |

Our method obtains the best performance compared to the other methods for *one vs. many* case. As one of the first attempts to achieve one-shot unpaired cross-domain image-to-image translation in the many-to-one setting, our work is confronted with more challenges, thereby possibly failing to get satisfactory results in some cases (e.g., generating various background) in Fig. 10 and Fig. 11.

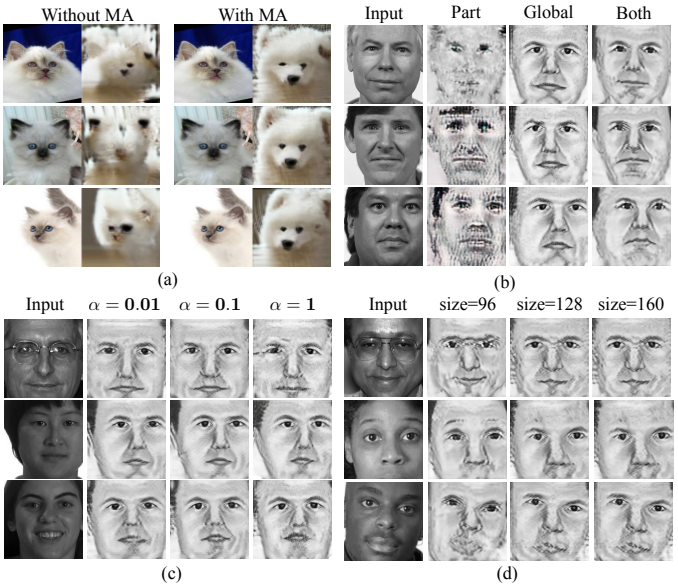



Fig. 12. The effectiveness of (a) multi-adversarial discriminators, (b) part discriminator and global discriminator, (c) the hyper-parameter $\alpha$ and (d) the size of cropped region.

### *E. Ablation study*

To investigate the efficacy of different components in our approach, we design several additional experiments for ablation study. Tab. VI shows the FID values of different variants of our approach on cat↔dog task. As can be seen, the performance drops dramatically by comparing GAN only with GAN+MA, which indicates that when using the multi-adversarial discriminators, the model can increase the ability to capture fine-grained semantic information. Fig. 12(a) also gives the visualized results of a similar conclusion. Fig. 3 exhibits image translation results with a local semantic matching while using additional part discriminators. We note that the translated images have a corresponding semantic link (mainly pose and layout) to their inputs.

The effectiveness of global discriminators and part discriminators can be shown in Fig. 12(b) and Tab. VI. It is easy to find that the translated images have very poor global consistency without global constraints if we do not use global discriminators. When we only use global discriminators, the generated images look similar to the same artifacts in the same region in all generated images. When using part discriminators, the performance can be improved, which benefits from the fact that the part discriminator can help alleviate the over-fitting by enforcing the generator to pay attention to the more fine-grained semantic details.

Additionally, we investigate the effects of the part size of part discriminators and hyper-parameter $\alpha$. Tab. VII shows the experimental results on PHOTO-SKETCH dataset and Fig. 12(c) exhibits visual results using different part sizes. Fig. 12(d) shows some examples by using different values of $\alpha$. We get the best performance when we choose the part size as $128 \times 128$ and $\alpha = 0.1$. For part learning, we also checked the different operations to obtain the part regions. For the photo-to-caricature image translation, face parsing guided crop can be also applied. The face parsing guided crop can make the part-discriminator easier to learn the semantic-level feature representations. But there are some drawbacks to using the face parsing guided crop: on the one hand, the correct face parsing label is difficult to obtain at the one-shot setting; on the other hand, the face parsing guided crop is only limited for face-to-face translation. By contrast, the random crop strategy can be applied for various translation tasks. We have also provided a comparison on using random crop and face parsing guided crop for one-shot photo-to-caricature translation, the quantitative comparison is shown in Tab. VIII. Our random cropping strategy has achieved better translation performance for exploring the possible correspondences between parts.

TABLE VI
FID SCORES OF ABLATION STUDY FOR CAT↔DOG TRANSLATION TASKS ON CAT2DOG DATASET.

| Method | dog→cat | cat→dog |
|---|---|---|
| GAN only | 398.6 | 413.3 |
| GAN + Balance | 336.4 | 285.1 |
| GAN + MA | 146.4 | 90.89 |
| GAN + Part + MA | 189.6 | 122.5 |
| GAN + Global + MA | 138.6 | 84.26 |
| GAN + Part + Global + MA | 132.3 | 73.26 |
| Proposed | **124.8** | **66.64** |

TABLE VII
QUANTITATIVE RESULTS OF USING DIFFERENT PART SIZES AND VALUES OF $\alpha$ FOR PHOTO→SKETCH TRANSLATION TASK ON PHOTO-SKETCH DATASET.

| Method | FID | LPIPS | SSIM |
|---|---|---|---|
| 96 (part size) | 116.3 | 0.4273 | 0.9026 |
| 128 (part size) | **94.28** | **0.3995** | **0.9172** |
| 160 (part size) | 105.0 | 0.4078 | 0.9030 |
| 0.01 ($\alpha$) | 116.3 | 0.4042 | 0.8961 |
| 0.1 ($\alpha$) | **94.28** | **0.3995** | **0.9172** |
| 1.0 ($\alpha$) | 137.6 | 0.4668 | 0.8953 |

TABLE VIII
QUANTITATIVE COMPARISON OF DIFFERENT SETTINGS FOR PHOTO-TO-CARICATURE IMAGE TRANSLATION TASK.

| Method | FID | LPIPS | SSIM |
|---|---|---|---|
| Random crop | **202.6** | **0.5523** | **0.9629** |
| Face parsing guided | 214.1 | 0.5718 | 0.9415 |

We have also implemented the experiments of sharing the backbone of the generators of both translation and the reconstruction functions. By sharing the parameters of the two generators, we can constitute a shared latent space between the source domain and the target domain. The domain-agnostic representations can boost the translation performance at the *photo-to-caricature* setting as shown in the Tab. IX. However, when the two domains have larger semantic distance at the *cat-to-dog* setting, the sharing parameters will harm the translation ability and lead to the worse results.

TABLE IX
QUANTITATIVE COMPARISON OF DIFFERENT SETTINGS. THE RESULTS ABOVE THE BOLD BLACK LINE REPRESENT THE RESULTS AT THE PHOTO-TO-CARICATURE SETTING, WHILE THE RESULTS BELOW THE BOLD BLACK LINE SHOW THE COMPARISON AT THE CAT-TO-DOG SETTING.

| Method | FID | LPIPS | SSIM |
|---|---|---|---|
| Proposed | 202.6 | **0.5523** | 0.9629 |
| Proposed (sharing) | **194.7** | 0.5616 | **0.9651** |
| Proposed | **64.64** | - | - |
| Proposed (sharing) | 92.89 | - | - |

### F. Limitation and failure cases

We evaluated our method on a variety of one-shot image-to-image translation tasks, and the results were not always satisfactory. We analyze the underlying reason causing such a phenomenon is the limitation of our approach for handling "unknown" objects. During the translation process, if the source object and the target object are not correlated, then the translation becomes difficult. For instance, in the cat↔dog task, though we can accomplish the translation of the main object from cat to dog, the background of the source image cannot be preserved. One possible reason is that "cat" and "dog" are semantically similar to each other, while the background is not necessarily correlated to "dog", thereby leading to the "abundance" of the background in the output image. We show more experimental results of failure cases on Cityscapes [71] and summer→winter [34]. Fig. 13(a) shows the failure results of the Cityscapes dataset. When the scene is complex and the only one image from the target domain could not cover all semantic information in the target domain on this semantic generation task. In Fig. 13(b), we conduct summer-to-winter translation while using one image depicting the winter scene. As can be observed, the translation fails to preserve "lake" in the generated image and the "cloud" is translated to "mountain" by mistake. Such phenomenons indicate that our approach tends to fail if the objects (e.g., "cloud", "lake") have never been observed in the target domain.

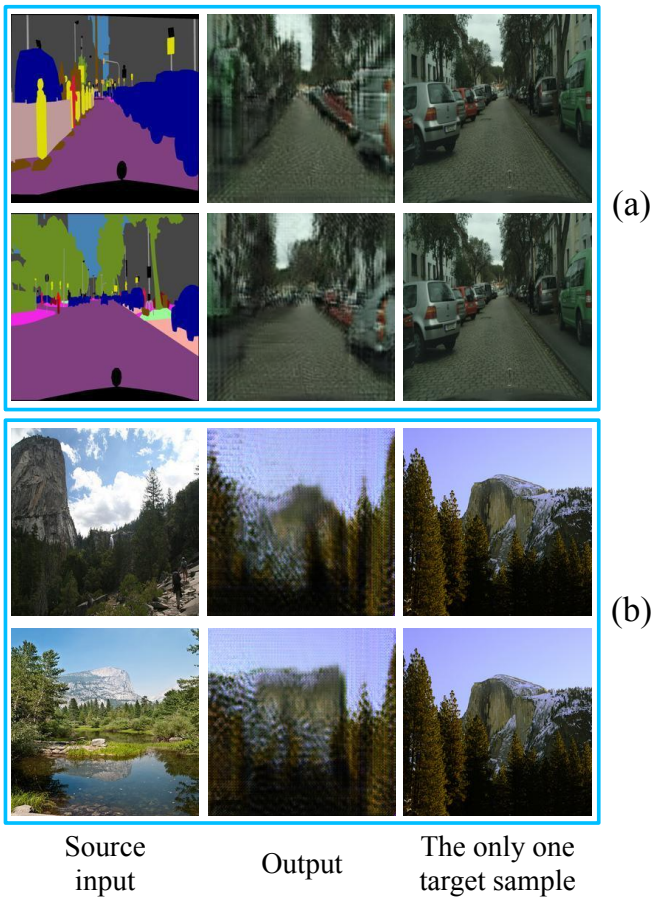



Fig. 13. Several failure cases of our method on Cityscapes and summer→winter datasets.

## V. CONCLUSION AND FUTURE WORK

In this paper, we proposed an effective method for one-shot cross-domain image-to-image translation to translate abundant samples from a source domain to another target domain with only one image. We introduced a multi-adversarial scheme to enhance the ability of discriminators to unearth effective information with given limited images. Besides, we included part-global learning architecture to extract more fine-grained information. Last but not least, we present a balanced adversarial loss to stabilize the adversarial training process and avoid over-fitting. We validated our method on multiple datasets and proved that our model can make use of the diversity information from the source domain and generate various kinds of images for the target domain even if the target domain only contains one training sample.

As one of the first attempts to achieve one-shot unpaired cross-domain image-to-image translation in many-to-one setting, our work is confronted with more challenges, thereby possibly failing to get satisfactory results in some cases (e.g., generating various background). We have been trying to share more prior knowledge (e.g., background, position) between the source domain and target domain, which practically helps to achieve more promising and reasonable results (e.g., generating various backgrounds). Besides, using an attention map to separate background and foreground during translation might be also helpful. We leave this as our future work.

## ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China under grant U20B2063, the Sichuan Science and Technology Program, China, under grant 2020YFS0057, National Key Research and Development Program of China under grant No. 2018AAA0102200, the Fundamental Research Funds for the Central Universities under Project ZYGX2019Z015, and Dongguan Songshan Lake Introduction Program of Leading Innovative and Entrepreneurial Talents.

## REFERENCES

[1] X. Shu, G.-J. Qi, J. Tang, and J. Wang, "Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *ACMMM*. ACM, 2015, pp. 35–44.
[2] W. Wang, G. Chen, A. T. T. Dinh, J. Gao, B. C. Ooi, K.-L. Tan, and S. Wang, "Singa: Putting deep learning in the hands of multimedia users," in *ACMMM*. ACM, 2015, pp. 25–34.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.

[7] S. Liao, X. Li, H. T. Shen, Y. Yang, and X. Du, "Tag features for geo-aware image classification," *IEEE transactions on multimedia*, vol. 17, no. 7, pp. 1058–1067, 2015.

[8] A. Potapov, I. Zhdanov, O. Scherbakov, N. Skorobogatko, H. Latapie, and E. Fenoglio, "Semantic image retrieval by uniting deep neural networks and cognitive architectures," in *AGI*. Springer, 2018, pp. 196–206.

[9] X. Ji, W. Wang, M. Zhang, and Y. Yang, "Cross-domain image retrieval with attention modeling," in *ACMMM*. ACM, 2017, pp. 1654–1662.

[10] P. Wu, S. C. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao, "Online multimodal deep similarity learning with application to image retrieval," in *ACMMM*. ACM, 2013, pp. 153–162.

[11] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 370–381, 2015.

[12] L.-W. Kang, C.-Y. Hsu, H.-W. Chen, C.-S. Lu, C.-Y. Lin, and S.-C. Pei, "Feature-based sparse representation for image similarity assessment," *IEEE Transactions on multimedia*, vol. 13, no. 5, pp. 1019–1030, 2011.

[13] Y. Peng and J. Qi, "Cm-gans: cross-modal generative adversarial networks for common representation learning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1, p. 22, 2019.

[14] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Transactions on Cybernetics*, vol. 50, no. 6, pp. 2400–2413, 2020.

[15] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 2017 ACM on Multimedia Conference*, 2017, pp. 154–162.

[16] Z. Wang, K. Chen, M. Zhang, P. He, Y. Wang, P. Zhu, and Y. Yang, "Multi-scale aggregation network for temporal action proposals," *Pattern Recognition Letters*, vol. 122, pp. 60–65, 2019.

[17] Y. Wang, L. Zhang, F. Nie, X. Li, Z. Chen, and F. Wang, "Wegan: Deep image hashing with weighted generative adversarial networks," *IEEE Transactions on Multimedia*, 2019.

[18] Y. Yao, F. Shen, J. Zhang, L. Liu, Z. Tang, and L. Shao, "Extracting multiple visual senses for web learning," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 184–196, 2018.

[19] W. Zhou, M. Yang, H. Li, X. Wang, Y. Lin, and Q. Tian, "Towards codebook-free: Scalable cascaded hashing for mobile image search," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 601–611, 2014.

[20] Y. Peng, J. Zhang, and Z. Ye, "Deep reinforcement learning for image hashing," *IEEE Transactions on Multimedia*, 2019.

[21] F. Shen, C. Shen, W. Liu, and H. T. Shen, "Supervised discrete hashing," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2015, pp. 37–45.

[22] H. T. Shen, L. Liu, Y. Yang, X. Xu, Z. Huang, F. Shen, and R. Hong, "Exploiting subspace relation in semantic labels for cross-modal hashing," *IEEE Transactions on Knowledge and Data Engineering*, p. 10.1109/TKDE.2020.297005, 2020.

[23] Y. Luo, Y. Yang, F. Shen, Z. Huang, P. Zhou, and H. T. Shen, "Robust discrete code modeling for supervised hashing," *Pattern Recognition*, vol. 75, pp. 128–135, 2018.

[24] N. Inoue and K. Shinoda, "Few-shot adaptation for multimedia semantic indexing," in *ACMMM*. ACM, 2018, pp. 1110–1118.

[25] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICMLW*, vol. 2, 2015.

[26] O. Vinyals, C. Blundell, T. P. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," *CoRR*, vol. abs/1606.04080, 2016. [Online]. Available: http://arxiv.org/abs/1606.04080

[27] Y. Duan, M. Andrychowicz, B. Stadie, O. J. Ho, J. Schneider, I. Sutskever, P. Abbeel, and W. Zaremba, "One-shot imitation learning," in *NeurIPS*, 2017, pp. 1087–1098.

[28] F. F. Li, R. VanRullen, C. Koch, and P. Perona, "Rapid natural scene categorization in the near absence of attention," *Proceedings of the National Academy of Sciences*, vol. 99, no. 14, pp. 9596–9601, 2002.

[29] E. G. Miller, N. E. Matsakis, and P. A. Viola, "Learning from one example through shared densities on transforms," in *CVPR*, vol. 1. IEEE, 2000, pp. 464–471.

[30] Q. Cai, Y. Pan, T. Yao, C. Yan, and T. Mei, "Memory matching networks for one-shot image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4080–4088.

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

[32] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*. Springer, 2016, pp. 694–711.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014, pp. 2672–2680.

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2242–2251.

[35] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *ICCV*, 2017, pp. 2868–2876.

[36] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, 2018, pp. 179–196.

[37] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *NeurIPS*, 2017, pp. 700–708.

[38] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *ECCV*, 2018, pp. 35–51.

[39] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 5967–5976.

[40] Z. Zheng, Z. Yu, H. Zheng, Y. Wu, B. Zheng, and P. Lin, "Generative adversarial network with multi-branch discriminator for cross-species image-to-image translation," *arXiv preprint arXiv:1901.10895*, 2019.

[41] S. Qiu, Y. Zhao, J. Jiao, Y. Wei, and S. Wei, "Referring image segmentation by generative adversarial learning," *IEEE Transactions on Multimedia*, 2019.

[42] Y. Guo, Q. Chen, J. Chen, Q. Wu, Q. Shi, and M. Tan, "Auto-embedding generative adversarial networks for high resolution image synthesis," *IEEE Transactions on Multimedia*, 2019.

[43] S. Benaim and L. Wolf, "One-shot unsupervised cross domain translation," in *NeurIPS*, 2018, pp. 2108–2118.

[44] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014. [Online]. Available: http://arxiv.org/abs/1411.1784

[45] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *CVPR*, 2018, pp. 8798–8807.

[46] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *ICML*, 2017, pp. 1857–1865.

[47] Z. Zheng, Y. Wu, X. Han, and J. Shi, "Forkgan: Seeing into the rainy night," in *The IEEE European Conference on Computer Vision (ECCV)*, August 2020.

[48] A. Gokaslan, V. Ramanujan, D. Ritchie, K. In Kim, and J. Tompkin, "Improving shape deformation in unsupervised image-to-image translation," in *ECCV*, 2018, pp. 649–665.

[49] J. Li, "Twin-GAN–unpaired cross-domain image translation with weight-sharing GANs," *arXiv preprint arXiv:1809.00946*, 2018.

[50] Y. Long and L. Shao, "Learning to recognise unseen classes by a few similes," in *ACMMM*. ACM, 2017, pp. 636–644.

[51] Z. Chen, Y. Fu, Y. Zhang, Y. Jiang, X. Xue, and L. Sigal, "Semantic feature augmentation in few-shot learning," *CoRR*, vol. abs/1804.05298, 2018. [Online]. Available: http://arxiv.org/abs/1804.05298

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

[53] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4570–4580.

[54] I. Durugkar, I. Gemp, and S. Mahadevan, "Generative multi-adversarial networks," in *ICLR*, 2017, pp. 1–14.

[55] E. Hosseini-Asl, Y. Zhou, C. Xiong, and R. Socher, "A multi-discriminator CycleGAN for unsupervised non-parallel speech domain adaptation," *Interspeech*, pp. 3758–3762, 2018.

[56] C. Hardy, E. L. Merrer, and B. Sericola, "MD-GAN: Multi-discriminator generative adversarial networks for distributed datasets," *arXiv preprint arXiv:1811.03850*, 2018.

[57] Y. Yang, Z. Ma, Y. Yang, F. Nie, and H. T. Shen, "Multitask spectral clustering by exploring intertask correlation," *IEEE transactions on cybernetics*, vol. 45, no. 5, pp. 1083–1094, 2015.

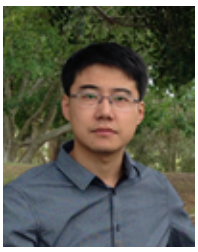
[58] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM TOG*, vol. 36, no. 4, pp. 1–14, 2017.
[59] E. Akleman, J. Palmer, and R. Logan, "Making extreme caricatures with a new interactive 2D deformation technique with simplicial complexes," in *Proceedings of Visual*, 2000, pp. 100–105.
[60] A. Mishra, S. N. Rai, A. Mishra, and C. Jawahar, "IIIT-CFW: A benchmark database of cartoon faces in the wild," in *ECCVW*, 2016, pp. 35–47.
[61] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015, pp. 3730–3738.
[62] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays, "Transient attributes for high-level understanding and editing of outdoor scenes," *ACM TOG*, vol. 33, no. 4, p. 149, 2014.
[63] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *CVPR*. IEEE, 2011, pp. 513–520.
[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[65] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *NeurIPS*, 2017, pp. 6626–6637.
[66] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are GANs created equal? a large-scale study," in *NeurIPS*, 2018, pp. 698–707.
[67] A. Borji, "Pros and cons of GAN evaluation measures," *CVIU*, vol. 179, pp. 41–65, 2019.
[68] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.
[69] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *PG*. IEEE, 2010, pp. 2366–2369.
[70] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE TPAMI*, vol. 31, no. 11, pp. 1955–1967, 2009.
[71] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016, pp. 3213–3223.

**Ziqiang Zheng** received his B.Eng. degree in communication engineering from Ocean University of China in 2019. He is currently a Master student in the College of Computer science and engineering at University of Electronic Science and Technology of China. His research interests include deep learning and computer vision.

**Zhibin Yu** (M'16) received the B.E. from Harbin Institute of Technology, China, the M.E. degree in computer engineering, and a Ph.D. degree in electrical engineering from Kyungpook National University, South Korea, in 2009 and 2016, respectively.

In 2016, he joined the Department of Electronic Engineering, Ocean University of China, where he is currently a Lecturer. His current research interests include underwater image processing, image-to-image translation, and generative neural networks.

**Haiyong Zheng** (M'12) received the B.Eng. degree in electronic information engineering and the Ph.D. degree in ocean information sensing and processing from the Ocean University of China, Qingdao, China, in 2004 and 2009, respectively. In 2009, he joined the Department of Electronic Engineering, Ocean University of China, where he is currently an Associate Professor. His research interests include image processing, computer vision, and machine learning.



**Yang Yang** is currently with the University of Electronic Science and Technology of China. He was a Research Fellow under the supervision of Prof. Tat-Seng Chua in the National University of Singapore during 2012-2014. He was conferred his Ph.D. Degree (2012) from The University of Queensland, Australia. During the Ph.D. study, Yang Yang was supervised by Prof. Heng Tao Shen and Prof. Xiaofang Zhou. He obtained Master Degree (2009) and Bachelor Degree (2006) from Peking University and Jilin University, respectively.

**Heng Tao Shen** is currently a Professor of National "Thousand Talents Plan", the Dean of School of Computer Science and Engineering, and the Director of Center for Future Media at the University of Electronic Science and Technology of China. He obtained his BSc with 1st class Honours and Ph.D. from the Department of Computer Science, National University of Singapore in 2000, and 2004 respectively. He then joined the University of Queensland as a Lecturer, Senior Lecturer, Reader, and became a Professor in late 2011. His research interests mainly include Multimedia Search, Computer Vision, Artificial Intelligence, and Big Data Management. Heng Tao has published 200+ papers, most of which appeared in prestigious publication venues of interests, such as ACM Multimedia, CVPR, AAAI, IJCAI, SIGMOD, VLDB, ICDE, TOIS, TIP, TPAMI, TKDE, VLDB Journal, etc. He has received 7 Best Paper Awards from international conferences, including the Best Paper Award from ACM Multimedia 2017 and Best Paper Award - Honorable Mention from ACM SIGIR 2017. He got the Chris Wallace Award for Outstanding Research Contribution in 2010 conferred by Computing Research and Education Association, Australasia, and the Future Fellowship from Australia Research Council in 2012. He has served as a PC Co-Chair for ACM Multimedia 2015 and currently is an Associate Editor of IEEE Transactions on Knowledge and Data Engineering. He is an Honorary Professor at the University of Queensland and holds a position of Visiting Professor at Nagoya University and the National University of Singapore.