Generative Adversarial Network with Multi-branch Discriminator for imbalanced Cross-species Image-to-image Translation

Ziqiang Zheng[a], Zhibin Yu[a], Yang Wu[b], Haiyong Zheng[a,*], Bing Zheng[a,], Minho Lee[c,*]

[a]No.238, Songling Road, Ocean University of China, Qingdao, Shandong, China
[b]Kyoto University, Kyoto, Japan
[c]Kynpook National University, Daegu, Korea

Abstract

There has been an increased interest in high-level image-to-image translation to achieve semantic matching. Through a powerful translation model, we can efficiently synthesize high-quality images with diverse appearances while retaining semantic matching. In this paper, we address an imbalanced learning problem using a *cross-species* image-to-image translation. We aim to perform the data augmentation through the image translation to boost the performance of imbalanced learning. It requires the model's strong ability to perform a biomorphic transformation on a semantic level. To tackle this, we propose a novel, simple, yet effective and efficient structure of Multi-Branch Discriminator (MBD) based on Generative Adversarial Networks (GANs). We show the effectiveness of the proposed MBD through theoretical analysis as well as empirical evaluation. We provide theoretical proof of why the proposed MBD is an effective and optimal case to have the best performance. Comprehensive experiments on various cross-species image translation tasks illustrate that our MBD can dramatically improve the performance of popular GANs with state-of-the-art results in terms of both objective and subjective assessments. Complete downstream image recognition evaluations at a few-shot setting have also been conducted to show that the proposed method can effectively boost the performance of imbalanced learning.

Keywords: Multi-branch Discriminator, Image-to-image translation, Generative adversarial network, Cross-species

*Corresponding author

1. Introduction

A common phenomenon in our daily life is that the examples from some species are significantly more than other species. Imbalanced datasets bring a challenge to most of the existing machine learning methods. Without intervention, many machine learning approaches tend to focus on the majority group while omit the minority groups. However, in many cases, the minority groups contain valuable information. Imbalanced datasets are very common in nature. For example, the number of pandas is much fewer than grizzly bear. Despite deep learning based approaches made a great progress in many computer vision fields[40, 78, 60, 14], the works on image-to-image based imbalanced learning are still quite limited.

The development of generative adversarial networks (GANs) brings new opportunities for the imbalanced learning challenge. GAN, which has been developed by Goodfellow*et al.*[21], is a proven as one powerful framework to handle various computer vision tasks [24], such as generating pictures from text descriptions [53, 76], converting video from still images [10], increasing resolution of images [36], editing and translating images/videos [75, 28, 81, 74, 66, 77, 63, 37, 27, 67]. A plausible GAN based solution for imbalanced learning is to take advantage of the diversity from rich species to generate reasonable samples for the rare categories to reduce the imbalanced ratio of the dataset. In particular, as an important and applicable topic in computer vision, GAN-based image-to-image translation has attracted more and more attentions [26]. Many extensions of the GAN [34] have focused on how to enhance the generation and synthesis ability to obtain better image translation performance by including new loss functions [3, 46], more complex architectures [79], as well as multiple networks [16, 81, 32, 72]. Some recent works [27, 37, 38, 20] started to address the issue of animal image-to-image translation tasks, such as cat↔dog and cat↔leopard. However, the mentioned translation mainly focuses on the pose matching of similar species while cannot work on further complex semantic matching situations within farther species (please see Fig. 1 for reference). The position, number and pose matching may bring additional diversity rather than conventional data augmentation (e.g. rotation, flipping or cropping).

We believe that the *cross-species* image-to-image translation with a semantic matching is a tough but meaningful work. In technical point of view, cross-species image-to-image translation tasks are extremely challenging, because such tasks requires powerful models to understand the semantic content representation of each image. For the imbalanced image datasets including at least one dominant species and one rare species, we can boost the recognition performance by translating images from the dominant species to the rare specie to perform **instance-level data augmentation**. To achieve this, we propose a novel *cross-species* image-to-image translation, and conduct a semantic matching between two or more species during the translation. Through the image translation, we can increase the diversity of the rare species even at the imbalanced setting. Specifically, we retain the pose, position, number and other semantic features during a cross-species translation to enrich the rare specie. Our goal is to handle the inter-species similarities and intra-species differences at the same time, which requires better semantic understanding. However, due to the high requirement of semantic mapping, most conventional GANs ineffectively and inefficiently handle cross-species image translation [81]. Current studies demonstrate that an ensemble discriminator architecture like multiple parallel discriminators or multi-scale discriminator, is helpful to GANs for improving and stabilizing the performance of image generation [16, 50, 22, 2] as well as translation [25, 66, 27]. In our work, we consider that the validity of the ensemble discriminator mainly benefits from the boosting mechanism [30], aiming to construct a strong learner based on many weak learners [47, 55]. Inspired by this idea, we propose to break a common strong discriminator into multiple smaller ones (branches) as weak learners, named Multi-Branch Discriminator (**MBD**, please check section. 3.1 for detail), for taking advantage of ensemble discriminator while reducing the complexity of architecture as well as computation. As shown in Fig. 2, a simple and powerful image-to-image translation model is designed for high-level image-to-image translation (*e.g.*, cross-species) and relief the imbalanced learning problem by translating images from a dominant specie to a rare specie.

Our contribution is three-fold: At first, we thoroughly investigate current different structures of GAN ensemble discriminator and present our novel MBD for boosting GANs. Second, we theoretically and empirically study that the MBD branch number should correlate with the total channel number while channel number per branch should not be too large or too small (the same for each discriminator), since too many channels make information redundant, while too less may lead to insufficient knowledge for translation. Interestingly, we find that multiple branches of MBD essentially bootstrap for task allocation on the semantic level during translation to tackle high-level (such as cross-species) image translation well even with limited training samples (**few-shot setting**), while increasing the diversity of the rare species that benefits the classification process on an imbal-

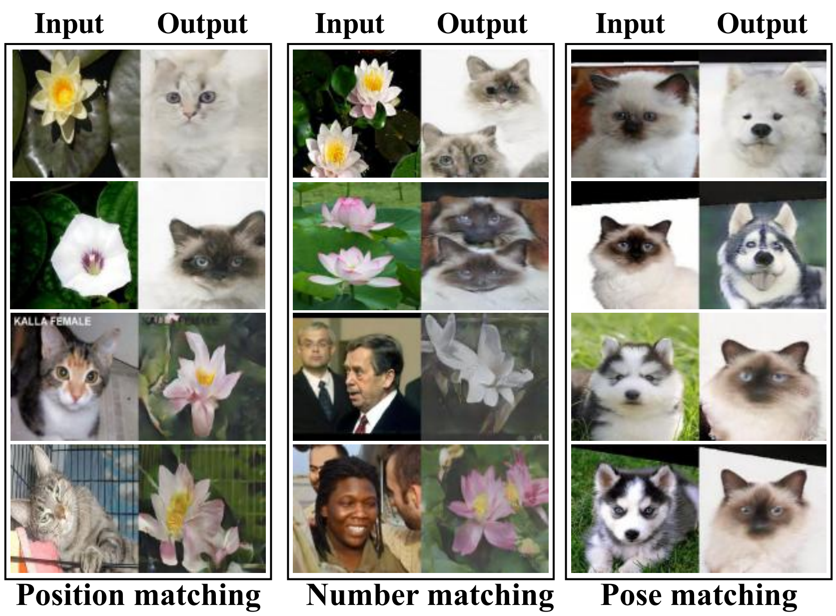


Fig. 1. Semantic level matching phenomenon including position matching, number matching and pose matching on *cross-species* image-to-image translation tasks by our proposed MBD method.

anced dataset.

2. Related works

2.1. Imbalanced learning

In past two decades, machine learning based imbalanced learning has been studied extensively. Methods can be crudely split into two groups: algorithm-level methods and data-level techniques. Algorithm-level methods try to address the imbalance problems by enhancing the importance of the minority categories on algorithm-level. Commonly used methods like modifying the cost functions, assigning penalties or weights for different categories try to force the model reduce the impact of imbalanced distribution. Wang et al. [65] presented Mean false error (MFE) for imbalanced classification tasks using deep neural networks. Lin et al. [39] proposed focal loss which can effectively addresses the extreme class imbalance in object detection tasks. Researchers soon found that the focal loss is also efficient for common image classification tasks [49]. Besides, there exists various kinds of cost-sensitive models such like CSDNN [64], CoSen- CNN [31], CSDBN-DE [73], etc. These models consider cost-sensitive learning to update

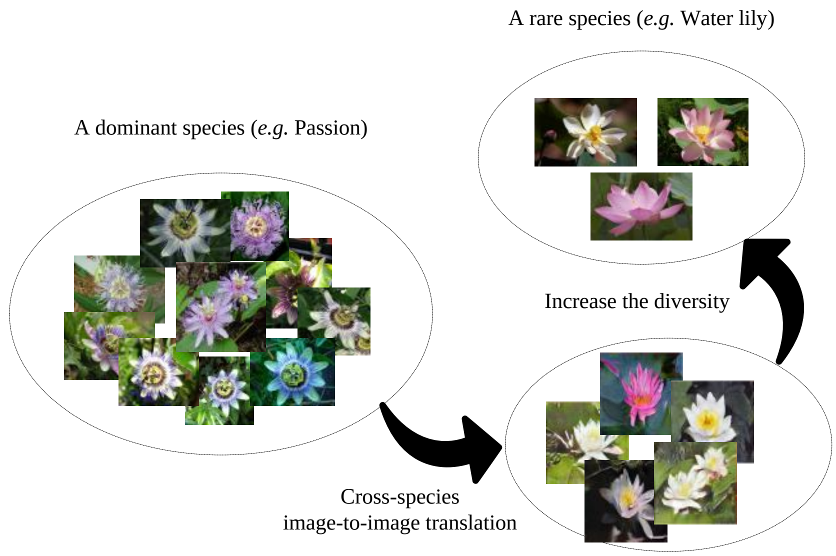


Fig. 2. *Cross-species* image-to-image translation for imbalanced learning.

model parameters. Generally, the cost-sensitive methods can significantly improve the imbalanced training process if a suitable cost/weight matrix is chosen. However, people may need experience to obtain an effective cost/weight matrix.

Instead of changing the algorithms, data-level techniques aim to modify the number of training data directly to reduce the imbalanced ratio. Over-sampling and under-sampling are two main strategies of data-level techniques. These methods try to reduce the imbalanced ratio among different categories on data-level by reducing the number of a majority category (e.g. random under-sampling) or increasing the number of a rare category (e.g. random over-sampling)[62]. Multiple data-level approaches are developed based on intelligent under/over-sampling [29]. However, some valuable information may also be discarded during the under-sampling process. The main weakness of over-sampling is that the increased samples will also increase the training time and has also been proved as a cause to over-fitting [11]. Recently, the development of generative models (such as GANs) brings a new direction to address the imbalance problems on data-level. Unlike over-sampling approaches, GANs based model can generate realistic image samples which are similar but different from the source samples. The adversarial training process make the generated samples robust against over-fitting. In this paper, we aim to develop a novel data-level framework for imbalanced learning based on generative adversarial learning.

2.2. Image-to-image translation

In general, image-to-image translation describes a task to convert an image of one source domain to an image in the target domain. Many typical computer vision topics can be summarized as image-to-image translation tasks [26], including semantic segmentation [42, 71], image restoration and enhancement [45, 75], image editing and in-painting [19, 52, 70, 5], super resolution [36, 12, 58]. In some early years, these tasks have been handled with various types of artificial neural network models [33, 54]. Due to the success of extensions on a conditional GAN [48], Isola *et al.* [28] developed an important branch of GAN called pix2pix to apply adversarial learning to image-to-image translation. Although pix2pix can handle general image-to-image tasks, it adopted a supervised training manner that always requires paired images. To overcome this shortage, Zhu *et al.* [81] proposed another variation called CycleGAN to extend GAN-based image-to-image translation to unpaired datasets with two generators (forward and backward). Soon after, Choi *et al.* [13] further extended this idea and proposed StarGAN to translate images among multiple domains with only one single generator and discriminator.

Along with the development of GAN techniques, many researchers have chosen unpaired training datasets for unpaired image-to-image translation [32, 72, 81, 82, 13, 27, 37]. However, cross-domain (*e.g.*, cross-species) translation tasks, are still considered to be extremely difficult [59, 43]. A recent study called MUNIT [27] adopted an unsupervised multi-modal structure to translate styles while preserving the contents to generate the target images. The concurrent DRIT [37] proposed a disentangled representation framework to generate diverse outputs with unpaired training data. Then, GANimorph [20] presented another unpaired image-to-image translation framework for shape deformation based on a discriminator with dilated convolutions. Besides, Twin-GAN [38] used a progressively growing skip connected encoder-generator structure for human-anime character translation. However, most of those works mainly addressed the pose matching situation of similar species, by using their specific-designed frameworks, which are difficult for reuse and recycle. In our work, we primarily target to take this one step further to study the challenging issue of *cross-species* image-to-image translation, which requires semantic level transformation. Also, we aim to build a simple and flexible yet effective structure based on current frameworks to gain the performance.

2.3. Ensemble discriminator GAN

Recently, many works have demonstrated that the ability of image generation and synthesis of GANs can benefit a lot from the design of ensemble discriminator, *i.e.*, multi-discriminator or multi-scale discriminator. GMAN [16] first extended GANs to multiple discriminators for high-quality image generation with fast and stable convergence. Multi-discriminator GAN (MD-GAN) [22] showed a learning procedure for GANs with multiple discriminators on the distributed datasets. MD-CycleGAN [25], which is an extension of CycleGAN [81], was proposed to enhance the speech domain adaption with an architecture of multiple and independent discriminators. Meanwhile, pix2pixHD [66] and MUNIT [27] adopted multi-scale discriminator structure for high-resolution paired and multi-modal unpaired image-to-image translation respectively. Besides, the works of Durugkar *et al.* [16], Doan *et al.*[15] and Albuquerque *et al.* [2] studied that the multiple discriminator setting can be helpful to stabilize GAN training.

We consider that the performance gain of ensemble discriminator GANs owes to the inside implicit boosting strategy. Boosting is an important branch of machine learning algorithms that construct a strong learner based on many weak learners [30, 55, 80]. For multi-discriminator GANs, the multiple and independent discriminators can be regarded as multiple weak learners, trying to construct a strong learner as well. In this paper, we empirically study the power of different types of multiple discriminators, which only have limited power on image-to-image translation tasks (please see Section 4.3.2 for reference). Unlike their methods, we present a novel ensemble discriminator framework by decomposing a common discriminator into multiple branches using channels as weak learners which are optimized independently (please see Fig. 3 for comparison). Comprehensive experiments demonstrate that this multi-branch discriminator outperforms the multi-discriminator structure on *cross-species* image-to-image translation tasks. It takes the advantage of ensemble discriminator while reducing the complexity of architecture and computation. There exists a boosting GAN, called AdaGAN [61], which is similar to AdaBoost [18]. It learns a weak GAN for each iteration concerning a re-weighted data distribution, rather we consider the boosting inside a discriminator of one GAN.

3. GAN-MBD

In this paper, we address one challenging *cross-species* image-to-image translation task. Our goal is to translate the images of one source species to target species to increase the diversity of the target category and relief the recognition

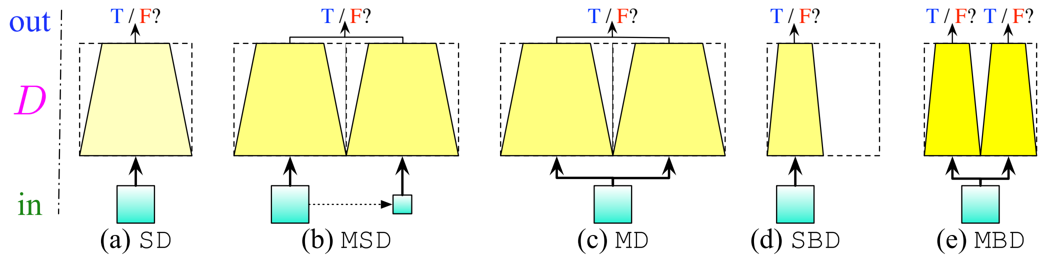


Fig. 3. Structure comparison of different types of GAN ensemble discriminator D. (a) SD: single discriminator [21]; (b) MSD: multi-scale discriminator [16]; (c) MD: multiple discriminators [27]; (d) SBD: single branch discriminator, *i.e.*, single discriminator with fewer channels; (e) MBD: our multi-branch discriminator. In this figure, for the "multiple" cases, MSD, MD and MBD, we present "2" for example illustration; for the SBD case, we present "$\frac{1}{2}$" channels for example illustration.

stress from an imbalanced dataset. Generally, the major challenge comes from the distance between species, *e.g.*, the task of flower↔human translation is harder than the task of cat↔dog translation. However, some other factors can also affect the translation difficulty, such as the position or number of species displayed in the image (please refer to Fig. 1). To meet the challenge, we develop efficient generative adversarial network of Multi-Branch Discriminator (GAN-MBD), which can handle the cross-species image-to-image tasks.

3.1. Multi-Branch Discriminator

Suppose a common discriminator has M channels for the ith layer, Our discriminator with N branches has M/N channels for each branch of the ith layer. Each branch that can be considered as a weak discriminator works independently. Notably, the number of parameters of a discriminator with two branches would only be half of a common discriminator (please refer to Table 1). Theoretically, we can obtain fewer parameters if we use more branches.

Fig. 3 shows the structure comparison of different types of the current GAN ensemble discriminator with our multi-branch discriminator. It can be seen that,

- Compared to the structure of multi-scale and multiple discriminator (MSD and MD), our multi-branch discriminator (MBD) structure is lightweight and easy to use. Besides, the optimization for our every branch of the discriminator is independent while the multiple discriminators are usually optimized together [66, 27, 22]. Some works [16, 50] also used independent optimization for multiple discriminators on image generation, but the generated images are restricted to low resolution and have obvious artifacts. Recent literature has addressed the issue of multi-discriminator training from the view of multiple random projections [50] and multi-objective optimization [2], further confirming that the multi-discriminator setting is helpful to stabilize and optimize GANs. Our study indicates that independent optimization for `MBD` is better than joint training and good enough for challenging *cross-species* image translation tasks (please see Section 4.3.2 for details).

- Compared to single discriminator with fewer channels (we describe this case as `SBD` that means single branch discriminator), our `MBD` can "understand" the discriminative task better, due to more information stored in more channels. This ensures each branch of discriminator is trained in charge of one sub-task being as a weak discriminator. Our detailed study in Section 4.3.2 also demonstrates that multiple branches of discriminator do act as weak discriminators for sub-tasks of translation and constitute one strong discriminator for the whole image-to-image translation.

The overall structure of our `MBD` takes advantage of both performance of multi-discriminator (`MSD` and `MD`) and lightweight of single-discriminator (`SD` and `SBD`). Please see Section 4.3.2 for detailed empirical comparison and analysis.

3.2. The variance of multi-branch regression

In this part, we will discuss the error variance between a multi-branch and a single-branch architecture. Let x and y denote the input and output images, respectively. The adversarial loss function of a common image-to-image translation model can be written as follows:

$$\begin{aligned} L(G,D) &= \mathbb{E}_{y\sim p_{data}(y)}\left[\log D\left(y\right)\right] \\ &+\mathbb{E}_{x\sim p_{data}(x)}\left[\log(1-D(x,G(x))\right] \end{aligned} \tag{1}$$

To handle both two-species and multi-species image-to-image translation, each branch should be able to distinguish two aspects. The first is to decide whether the current image is real or synthesized. Other parallel task is to judge the image category. The average output of all branches is the final result of the discriminator. Thus, the loss function of `MBD` model is:

$$\begin{aligned} \mathcal{L}(G,D) =& \frac{1}{N}\sum_k^N \{\mathbb{E}_{y\sim P_{\text{data}}(y)}\left[\log D_k(y)\right]+ \\ & \mathbb{E}_{x\sim P_{\text{data}}(x)}\left[\log(1-D_k(G(x))\right]\} + \mathcal{L}_c, \end{aligned} \tag{2}$$

where x denotes the original real sample, y means the target real sample, D_i is the ith branch of discriminator D, N is the number of branches, and $\mathcal{L}_c$ is the category classification loss which can be written below:

$$\begin{aligned}\mathcal{L}_c = &\frac{1}{N}\sum_i^N \{\mathbb{E}_{y,c}[-\log D_{c,i}(c|y)]+ \\ &\mathbb{E}_{x,c}[-\log D_{c,i}(c|G(x))]\},\end{aligned} \tag{3}$$

where the first term and the second term denote the classification loss for real samples and fake samples respectively, c represents the label, and $D_{c,i}$ means the category identification task for the ith branch.

Let e_i be the estimation error of the branch i. We assume $e_i \sim N(\mu, \sigma^2)$ is Gaussian distributed and the correlation coefficient between the output of branch i and j is ρ_{ij}. Suppose each branch has the same power, we can obtain the variance of a multi branch architecture as follows:

Proposition 3.1. *The variance of multi branch estimation* $\mathrm{Var}(\frac{1}{N}\sum_{i=1}^N e_i)$ *has the following property:*

$$\frac{1}{N}\sigma^2 \leq \mathrm{Var}(\frac{1}{N}\sum_{i=1}^N e_i) \leq \sigma^2$$

Proof 3.1.

$$\begin{aligned}
\mathrm{Var}(\frac{1}{N}\sum_{i=1}^N e_i) &= \mathrm{Var}(\sum_{i=1}^N \frac{1}{N}e_i) && (4)\\
&= \mathrm{Cov}(\sum_{i=1}^N \frac{1}{N}e_i, \sum_{i=1}^N \frac{1}{N}e_j) && (5)\\
&= \sum_{i=1}^N (\frac{1}{N})^2 \mathrm{Var}(e_i) + 2\sum_{i,j=1;j\neq i}^N \rho_{ij}\frac{1}{N}\sqrt{\mathrm{Var}(e_i)}\frac{1}{N}\sqrt{\mathrm{Var}(e_j)} && (6)\\
&= \frac{1}{N}\sigma^2 + 2\sum_{i,j=1;j\neq i}^N \rho_{ij}\frac{1}{N^2}\sigma^2 && (7)\\
&= \frac{1}{N}\sigma^2 + N(N-1)\rho_{ij}\frac{1}{N^2}\sigma^2 && (8)\\
&= \frac{1}{N}\sigma^2 + \rho_{ij}(1-\frac{1}{N})\sigma^2 && (9)
\end{aligned}$$

Because $0 \le \rho_{ij} \le 1$*, we can derive*

$$\frac{1}{N}\sigma^2 \le \mathrm{Var}(\frac{1}{N}\sum_{i=1}^{N} e_i) \le \sigma^2.$$

We can achieve the minimum error $\frac{1}{N}\sigma^2$ while $\rho_{ij} = 0$, which means that each branch plays **independently**. In this case, larger N implies smaller error. On the contrary, we will obtain the maximum error σ^2 if $\rho_{ij} = 1$, which means all the branches are in perfect correlation. In this situation, there will be no difference between N branches and one branch.

4. Experimental comparison

4.1. Datasets

Cat2dog [37] includes 871 cat and 1364 dog cropped images in total. We inherit this dataset following the same data split for training and testing with the ratio of 771:100 for cat and 1264:100 for dog, respectively.

102Flowers [51] contains 102 different categories of flowers. We choose five categories: grape hyacinth, water lily, rose, thorn apple, and hibiscus, with 704 images for training and 174 images for testing.

CelebA [41] is a large-scale face attributes dataset with more than 200K images. In our experiments, to focus on the face while translation, we randomly select and crop 801 facial images, and split with 695 for training and 106 for testing.

Dogs vs. Cats | Kaggle [17] includes 25,000 dog and cat images captured in the wild, which is more challenging than Cat2dog. We randomly select 627 cat images from this dataset with 526 and 101 for training and testing respectively, to explore the potential of our method on image translation in the wild.

LFW [35] contains more than 13,000 images with labeled faces in the wild. Compared to CelebA, LFW possesses more poses under more complex conditions, such as two people in one image. We randomly choose 1002 images from this dataset for the image translation in the wild, among which 804 for training and 198 for testing.

ODIR5K [1] contains fundus photographs of both the left and right eyes from 5000 patients. There are eight categories for all the samples, which includes normal, diabetes, glaucoma, cataract, AMD, hypertension, myopia and other diseases.

4.2. Evaluation metrics

Fréchet Inception Distance (FID) [23] is proposed to compute the distance between the generated sample distribution and real distribution. This method is a consistent and robust approach for evaluating the generated images [44, 8], which can be calculated by:

$$\text{FID} = ||\mu_x - \mu_g||_2^2 + \text{Tr}\left(\sum_x + \sum_g - 2(\sum_x \sum_g)^{\frac{1}{2}}\right), \tag{10}$$

where $(\mu_x, \sum_x)$ and $(\mu_g, \sum_g)$ are mean and covariance of the sample embeddings from the data distribution and model distribution. A lower FID score indicates higher generated image quality. We use FID as the main objective assessment of our experiments.

User study is still the golden standard for assessing the quality of generated images, especially for image translation, since it requires some kinds of semantic mapping that are hard to be calculated [6, 56]. To evaluate the image translation quality, referring to [7], we ask 20 persons to rate whether the target image matches the source image (presenting the methods and samples in random order), and calculate the ratio of "yes" answers as the grade. We use user study as a subjective assessment for our experiments.

Classification accuracy is also combined. To demonstrate that the proposed method can boost the performance of downstream recognition task, we adopt this metric to evaluate the classification performance, and the higher the better.

4.3. Ablation study for MBD

4.3.1. How many branches and channels are better?

Since our `MBD` decomposes a discriminator into branches by channels, so the branch number N should correlate with the total channel number M, but how to choose the number of channels of each branch for better performance, and how many branches are reasonable? In the previous section, we have already proved that we can obtain the minimum error when we have N independent branches. Ideally, the larger N we have the lower error we can get. But it would become increasingly difficult to make all the branches independent branches as N grows. Also, total number of channels M increases the computational cost, making the optimization more difficult. Insufficient channels may not have enough power to handle a classification task. Thus, for a general task, both M and N should not be too large or too small.

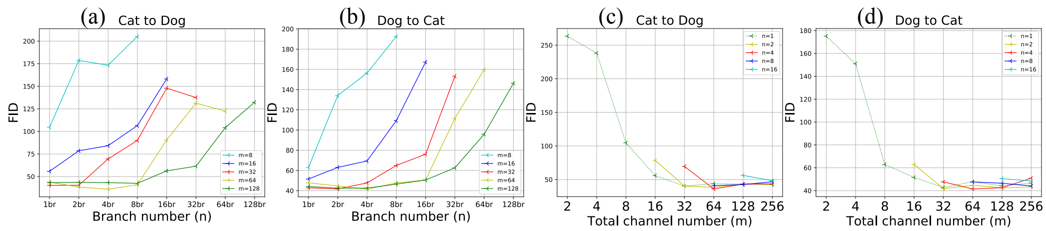


Fig. 4. The relation between branch number N and and total channel number M in terms of FID on cat$\leftrightarrow$dog translation, indicating that $M = 64, N = 4$ could be an optimal setting for further experiments.

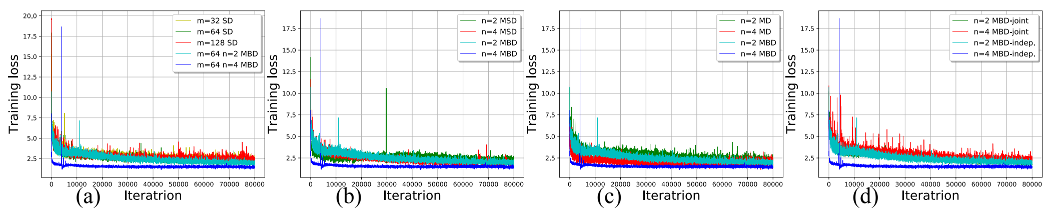


Fig. 5. Training losses of different ensemble discriminator structures, showing that our `MBD` (blue) can accelerate and stabilize convergence.

To find a suitable parameter combination for M and N for real *cross-species* tasks, we empirically study the relation between branch number N and total channel number M of a discriminator in terms of Fréchet Inception Distance(FID) metric (lower is better). We use cat$\leftrightarrow$dog on Cat2dog dataset in cross-species image translation task and adopt CycleGAN [81] as base architecture. As shown in Fig. 4 (a) (cat$\rightarrow$dog) and (b) (cat$\leftarrow$dog), most of the M curves indicate that the translation results would be worse (higher FID) as N gets bigger, which implies that too many branches may not be good for the performance. Besides, the $M = 8$ and $M = 16$ curves are higher than the other curves, which shows that the total number of channels should not be too small to feed enough information to the discriminator. From the three curves of $M = 32, 64, 128$, we can find that $M = 128$ performs not better than $M = 64$ and $M = 32$, indicating that too many channels may not be helpful, further, $N = 1$ cannot achieve the best results, that is, multi-branch discriminator, especially 2 and 4 branches, does work better than a common discriminator.

To further confirm the better branch number with a channel number, we change the view from N to M, as shown in Fig. 4 (c) (cat$\rightarrow$dog) and (d) (cat$\leftarrow$dog), almost each curve goes down and then goes up, demonstrating that the channel number per branch M/N also should not be too large or too small, indicating that there exists the optimal match point between N and M, which is $N = 4, M =$

64 at the lowest point. So we use this as the optimal setting for the following experiments, ***i.e.***, 4-branch discriminator with 16 channels per branch (4MBD).

Note that the $N = 1$ curves in Figs. 4 (c) and (d) can also show that the channel number should not be too large or too small for a common discriminator, beyond that, they actually compare our multi-branch case to the case of single discriminator with fewer channels (SBD in Fig. 3) if we choose the same M/N (*e.g.*, $M = 8$ at $N = 1$ and $M = 16$ at $N = 2$), demonstrating that MBD performs better than SBD. This conclusion can be verified further from Fig. 7. Therefore, we suggest that each branch/discriminator to have no less than 16 channels and no more than 128 channels to handle a common image-to-image task.

4.3.2. How good is our MBD?

Apart from for the structure comparison shown in Fig.3, we also use cat↔dog translation task on Cat2dog dataset for the empirical study of the different structures including SD, MSD, MD and our MBD. We build all the structures based on CycleGAN [81], with the total channel number $M = 64$ and the branch/discriminator number $N = 2, 4$ (we set M and N according to the performance shown in Fig.4).

Table 1 lists the FID and user study results of different structures with their discriminator parameter amounts on cat↔dog image translation task. It can be seen that, MSD performs even worse than SD, MD is a little superior to SD, while our MBD with 4 branches achieves the best FID and user scores with fewest parameters(Table 1). We also notice that 4MD has better user scores but not good FID results, and the reason may be the low diversity of images synthesized by 4MD. The visual results in Fig. 6 can also conclude that our MBD does outperform the single discriminator structure SD as well as multi-discriminator structures MSD and MD. Dog and cat images obtained by SD have unclear outlines. Some dogs and cats generated by 2MSD or 4MSD have abnormal ears and nose.

Table 1: FID and user study with discriminator parameter amounts (DParams, millions) comparison of different models on cat↔dog image translation.

Method	Cat→Dog		Dog→Cat		DParams
	FID	User	FID	User	
SD	43.8814	0.404	47.7190	0.676	13.92
2MSD	143.6501	0.181	86.4939	0.320	27.84
4MSD	56.5513	0.550	49.2084	0.750	55.68
2MD	43.6576	0.632	42.5042	0.905	27.84
4MD	43.6651	0.844	50.9160	0.862	55.68
2MBD	38.4135	0.683	44.7008	0.901	6.97
4MBD	**36.0023**	**0.850**	**41.4614**	**0.940**	**3.50**

Fig. 6. The cat→dog (left) and dog→cat (right) translation results of different ensemble discriminator structures.

We further study the training loss and parameters of different structures. As shown in Figs. 5 (a), (b) and (c), compares to SD, MSD and MD structures respectively, which shows our MBD (blue color) can accelerate and stabilize the convergence. Besides, Fig. 5 (d) illustrates that, for ensemble discriminator structure, independent optimization does work better than joint training. Moreover, the parameters of different GAN discriminators listed in Table 1 indicate that the MBD structure dramatically reduces the model complexity compared to other structures. Thus, we can say that MBD does render the optimization easier with lower training loss and less number parameters.

Finally, by visualizing feature maps of different layers in each branch of the discriminator, we surprisingly find that MBD essentially bootstraps for task allocation on the semantic level during translation. Brighter pixels mean higher activation response in heat maps [69] as shown in Fig. 8. We can find that the four branches of discriminator are learned for different sub-tasks automatically, *i.e.*, for the input cat image (left), the 1_{st} branch is responsible for contours, the 2_{nd} branch takes in charge of furs, the 3_{rd} branch is interested in eyes and edges, and the 4_{th} branch captures whiskers. Please note that the cat images and dog images come from the two different discriminators of CycleGAN for the two different domains (cat/dog). Thus, the heat maps of the four branches on the right seem to be a little different: the 1_{st} branch catches the mouth and the illumination, the heat maps of the 2_{nd} branch display grounds and furs, the 3_{rd} branch records contours, and the maps of the 4_{th} branch include eyes and whiskers.

As we mentioned in Section 3.2, our system would have the best performance if each branch works independently. Through these visual experiments, we can easily find that each branch of MBD has a clear division of labor. We believe that this fine labor division is a key to tackle high-level (*e.g.*, keeping pose matching during a cross-species task) image translation effectively.

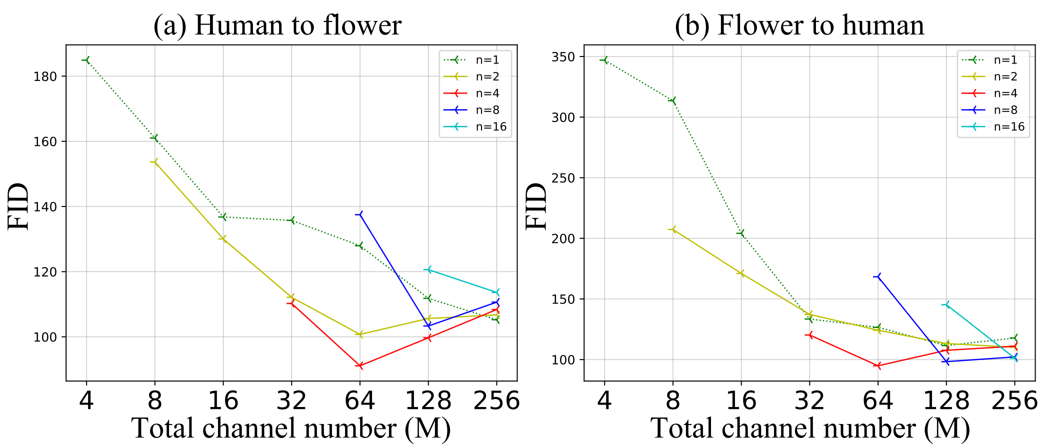


Fig. 7. The relation between branch number N and and total channel number M in terms of FID on more challenging flower↔human translation, further demonstrating that both `SD` and `SBD` ($N = 1$) perform worse than our `MBD`.

4.3.3. Why MBD works for imbalanced learning?

An imbalanced dataset contains at least a rare category A and a rich category B. Suppose the rich category B has sufficient diversity including different poses, numbers, positions and so on, while the rare category A has insufficient diversity due to the lack of samples. To make a decision boundary between the A and

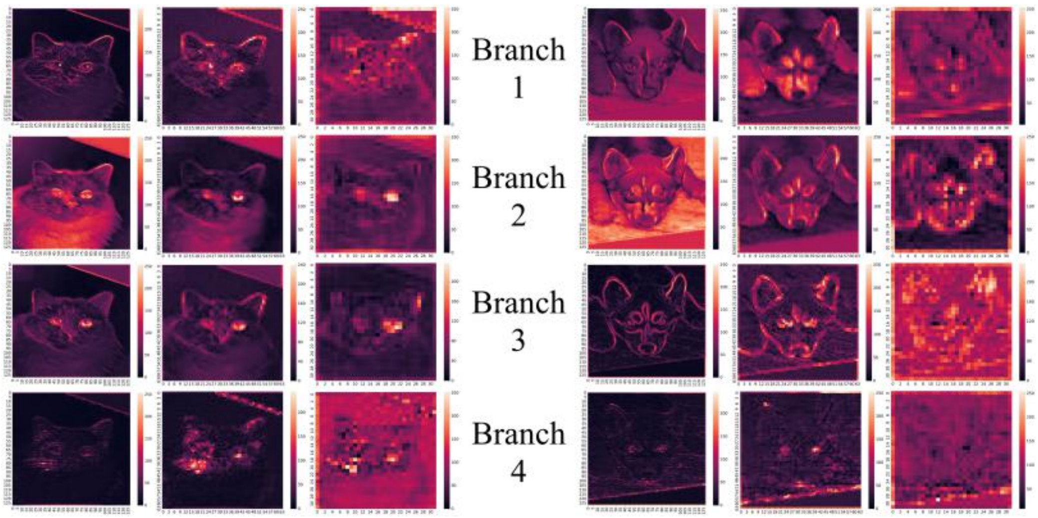


Fig. 8. Visualization of feature maps from `MBD` different layers shows bootstrapping task allocation of branches.

B, a baseline solution is to draw a boundary in the middle of the two datasets (Fig. 9(a)). However, due to lack of samples of A, the boundary cannot truly reflect the distribution of the two categories.

To improve the shape of the decision boundary, data-level machine learning methods tried to adjust the imbalance ratio using various under-sampling and over-sampling approaches (Figs. 9(b) and (d)) [62, 9]. Algorithm-level techniques commonly used a punishment/gain to the rich/rare category B/A (Fig. 9(c)) [39, 31]. Indeed, we can obtain a better decision boundary with those approaches than a baseline solution. But the decision boundary may be still inaccurate due to the lack of samples of the rare category A near the boundary. Inspired by the transfer learning strategy [68], if we can learn and translate the diversity of B to A following the semantic matching, we can draw a more reasonable decision boundary to improve the imbalanced classification accuracy (Fig. 9(e)). In other words, we can have a better chance to solve the imbalanced classification task if we can take advantages of the diversities on positions, poses, numbers and other cases from the rich category B to enrich the category A.

As demonstrated in Fig. 8 in section4.3.2, each branch of our MBD can focus on a specific translation task (e.g. eyes or whiskers) to tackle a semantic-level translation task like pose matching or number matching. The semantic-level image-to-image translation can practically help us to build a better decision boundary by generating reasonable and meaningful samples for the rare category

A.

Fig. 9. (a) Decision boundary of a baseline solution, (b) decision boundary with down sampling, (c) decision boundary with cost-sensitive loss function, (d) decision boundary with over sampling, (e) decision boundary with semantic-level augmentation.

4.4. Cross-species image translation

For image-to-image translation between two species, we use CycleGAN [81] as baseline to build **CycleGAN-4MBD** with 4-branch discriminator as our MBD model. We compare our method with two state-of-the-art methods: MUNIT [27] and DRIT [37]. And all the image results are listed with "input-output" pairs. We choose four species: cat, dog, flower and human, for our *cross-species* image-to-image translation experiments, due to the limited space, we illustrate cat↔dog, cat↔flower and flower↔human tasks in the paper, and leave other three tasks (cat↔human, dog↔flower and dog↔human) in the supplementary file. **Cat⇔Dog.** As mentioned in the literature [81, 27, 37], unpaired image-to-image translation between a cat and a dog is an open puzzle. We adopted the Cat2dog [37] dataset for this task, and the translation results are shown in Fig. 10. We observe that the images generated by our CycleGAN-4MBD exhibit the best performance. Notably, the synthesized dogs/cats of our method still maintain the same poses like those of the input images. Table 2 shows the FID and user study results of the three models, where our CycleGAN-4MBD also achieves the highest scores on both image translation tasks.

Table 2: FID and user study on cat↔dog image translation.

Method	Cat→Dog		Dog→Cat	
	FID	User	FID	User
DRIT	88.6275	0.138	58.4392	0.610
MUNIT	47.4142	0.510	46.2864	0.810
CycleGAN-4MBD	**36.0023**	**0.850**	**41.4614**	**0.940**

Cat⇔Flower. Another experiment is implemented between Cat2dog and 102Flowers datasets. We only choose cat images for cat↔flower translation. Compared with image translation between cats and dogs, flowers seemly do not have explicit pose or facial representation. Thus, it is not possible to expect a semantic pose matching between cats and flowers. However, the translation results shown on the left of Fig. 11 still demonstrate that our CycleGAN-4MBD can obtain position matching between these two farther species.

Fig. 10. The cat↔dog image translation results of our CycleGAN-4MBD compared to MUNIT and DRIT.

To explore the potential of our method, we implement a more challenging experiment based on Dogs vs. Cats | Kaggle and 102Flowers datasets, because both are captured in the wild. The translation results are displayed in the right of Fig. 11. Compared with above experiments, there is an additional number matching test from cat to flower due to the multiple instances. Also, the comparison illustrates that both MUNIT and DRIT may work when input image includes a single individual only, whereas, MUNIT fails to convert two flowers to two cats while our method can still do so. Besides, our method synthesizes the head of a cat from a flower but without the full-body, which makes a lot more sense. The FID results listed in Table 3 also confirm the best performance of our MBD on cat↔flower image translation.

Flower⇔Human. We also evaluate image translation between 102Flowers and CelebA datasets. Fig. 12 (left) shows that all 3 methods can accomplish a human→flower translation, however, the flower→human translation seems more challenging, where images synthesized by MUNIT or DRIT always have a serious distortion while our method can still achieve a reasonable solution.

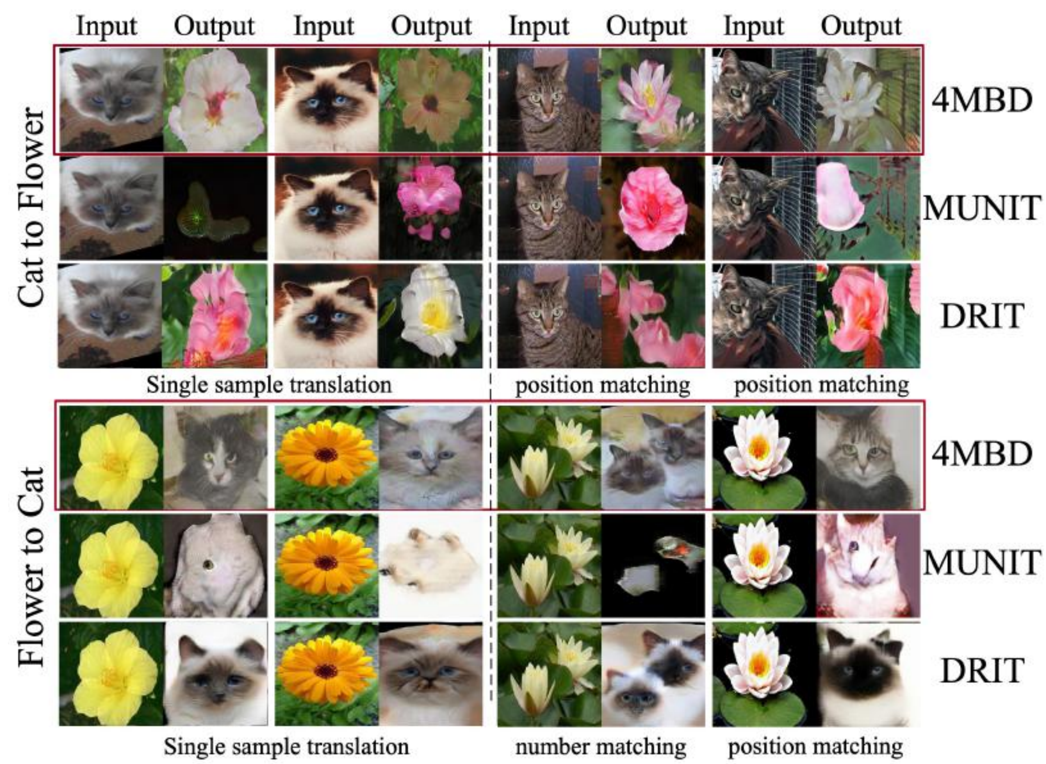


Fig. 11. The controlled (left) and wild (right) cat↔flower image translation results of our CycleGAN-4MBD compared to MUNIT and DRIT.

Table 3: FID on cat↔flower image translation.

Method	Cat→Flower		Flower→Cat	
	controlled	wild	controlled	wild
DRIT	183.0757	166.9990	157.1238	221.0960
MUNIT	169.9913	143.5711	141.0376	187.7106
CycleGAN-4MBD	**117.6854**	**125.9365**	**46.2705**	**128.1857**

We then designed another challenging experiment based on LFW and 102Flowers datasets, as both were captured in the wild. As shown in the right of Fig. 12, our method can handle most of human→flower cases. Although we find some unreasonable distortion, we can still observe position and number matching in wild flower→human translation. Table 4 also illustrates that our method can obtain the best FID results against MUNIT and DRIT.

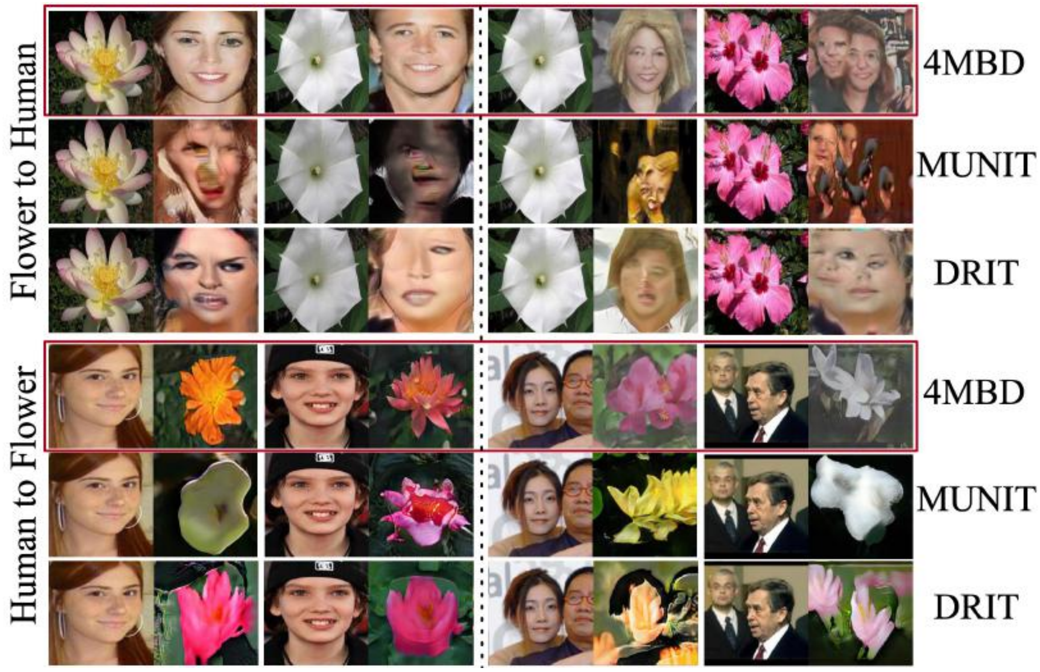


Fig. 12. The controlled (left) and wild (right) flower↔human image translation results of our CycleGAN-4MBD compared to MUNIT and DRIT.

x

Table 4: FID on flower↔human image translation.

Method	Flower→Human controlled	Flower→Human wild	Human→Flower controlled	Human→Flower wild
DRIT	142.2868	149.5487	165.8483	111.9950
MUNIT	241.9420	205.8153	147.6277	127.9194
CycleGAN-4MBD	**91.1127**	**136.0121**	**94.8746**	**102.8981**

4.5. Imbalanced cross-species image translation

4.5.1. Semantic-level translation

In this section, we perform various image-to-image translation at the imbalanced setting (the images from different domains are not equal). We have one target domain, which has only 10 samples (**few-shot** setting), while the source domain has redundant samples. Firstly, our method achieved the pose matching for the imbalanced Sword lily→Watercress image translation. There are 130 sword lily image and only 10 watercress images at the training stage. We observed the **pose**, **position** and **number** matching at the translation procedure as shown in Fig. 13. Our method can capture the pose expression of the input image and preserve the content information after the translation. The proposed method has a strong ability to extract the number information of the input images and translate the small flowers to the required representation.

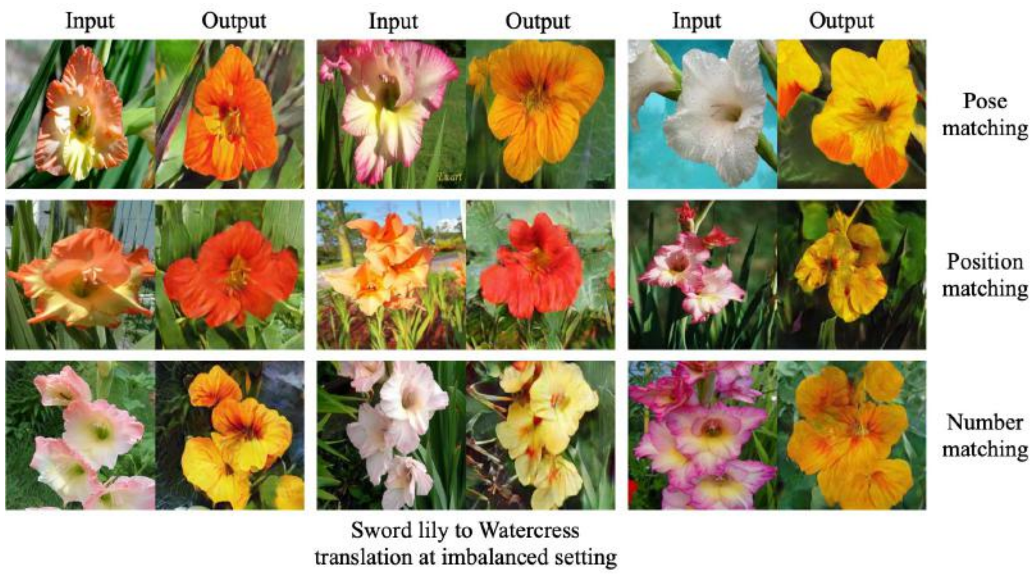


Fig. 13. The visual translation results of proposed method at imbalanced setting, the rare category "Watercress" has only ten images. Our method can achieve the pose, position and number matching during the translation procedure.

4.5.2. Data augmentation at imbalanced setting

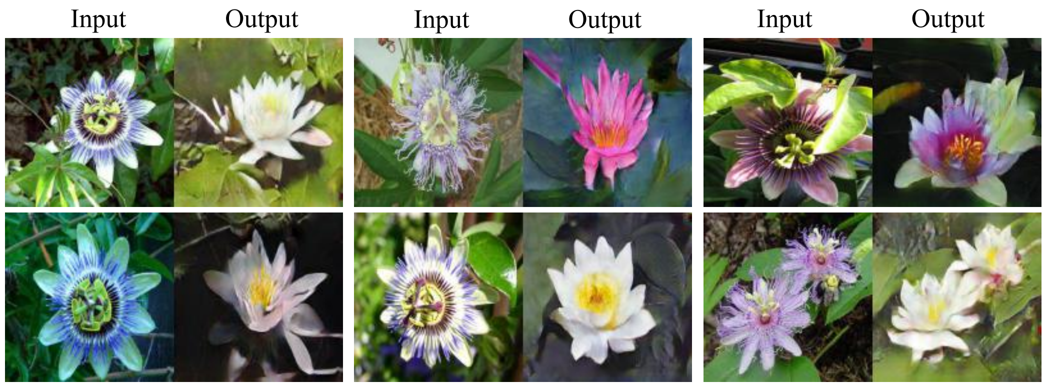
In this section, we discuss the potential applications of the proposal at the imbalanced setting. Sometimes it is difficult and time-consuming to collect a large number of images from a specified rare category. In many cases, insufficient samples often signifies poor diversity and effective information, which leads to that it is a huge challenge to perform image recognition based on few samples (e.g. 10 samples). For this problem, we can take advantage of the diversity from a dominant category with redundant samples and adopt the *cross-species* techniques to synthesize samples for the rare category, we can efficiently reduce the imbalanced ratio without using the cost-sensitive losses [39, 4]. We apply our method for the imbalanced cross-species image translation to achieve data augmentation.

We aim to translate the redundant images from the dominant category to the target rare category, which usually has quite few image samples. Through this way, we can obtain more translated fake images in the dominant category with semantic matching. We can obtain one better image classifier with the synthesized images. To prove our assumption, we perform data augmentation based on these translated images and boost the performance of image recognition in Table 5. The visual translated images are shown in Fig. 14. For image classification, we adopt the VGG-19 network[57] as our binary classifier and perform experiments at two settings: 1) 200 passion (dominant category) and 10 water lily (rare category) flower images; 2) 200 passion and 210 (200 translated images generated by our CycleGAN-4MBD and 10 original images) water lily flower images. For both the two settings, the validation set includes 51 passion and 51 water lily images, respectively. We obtain 66.67 percents accuracy on the validation set without any support at the first setting. Due to the redundant passion and insufficient water lily images, all the 51 passion images are successfully recognized. but 34 water lily images are wrongly classified as the passion category. Besides, we also adopt the focal loss [39] using the default parameters ($gamma = 2.0$) following the same train/test split. The focal loss is believed as a state of the art on imbalanced learning[29] and achieves 71.57 percents accuracy. By combining the proposed MBD, we can achieve a competitive 88.23% accuracy, which outperforms other methods a large margin. Following the same setting, we also perform experiments using 200 passion and 10 rose images. The vanilla model obtains about 64.7 percentage points. The focal loss brings about 11 percentage improvement. Our method can significantly enhance the average accuracy by more than 20 percents.

Table 5: The classification accuracy of the image classifier model at both two cases: with and without translation. With the translated images from the dominant category, the classification accuracy has been improved a lot.

Method	Data split		Classification accuracy		
	Source split	Target split	Cross entropy	Focal loss[39]	Cross entropy with translation
Passion and Water lily	200/10	51/51	66.67%(51/17)	71.57%(46/23)	88.23%(47/43)
Passion and Rose	200/10	51/51	64.70%(50/16)	76.47%(46/26)	89.21%(46/45)

Moreover, we also conduct the imbalanced data augmentation at a more challenging setting: we have one dominant category with redundant images and several rare categories with few images. We select 6 flower categories from 102Flowers dataset, which contain Passion flower (251 images), Water Lily (194 images), Rose (171 images), Windflower (54 images), English marigold (65 images) and tree poppy (62 images). At this setting, we use 200 Passion flower images for training, while only randomly choose 10 training images for other flower categories. Following the above experimental setting, the train/test split of the 6 categories are listed in Table 6. To be noted, we only use 44 Windflower images due to there are 54 images in total. The average classification accuracy of the vanilla



(a) Passion to Water Lily translation at imbalanced setting

(b) Passion and Rose translation at imbalanced setting

Fig. 14. The visual translation results of proposed method at imbalanced setting, the rare category "Water lily" or "Rose" has only ten images. We use our model to translate the redundant source images (Passion images) from the dominant category to the required Water lily images and Rose images. During the translation procedure, our method can achieve the position and number matching after the translation. More usefully, our method can even capture the pose representation and generate corresponding outputs with same pose.

model is 51.17%. Considering the traditional image processing methods: the flipping and randomly cropping can also be applied for data augmentation, we also perform experiments by using these two operations to perform up-sampling: we randomly flip the image and resize it to 256×256, and then randomly crop the resized image to 224×224 to obtain more training samples. We obtain 200 augmented samples based on the raw 10 images, which indicates that we obtain 20 different random augmented samples from one original sample. The experimental result comes to 52.84%. The focal loss can also promotes the classification accuracy at this setting (from 51.17% to 70.23%). Besides, to show that our *MBD* architecture really achieves the boost at the imbalanced setting, we apply the original CycleGAN to perform the translation among the categories to achieve data augmentation. The original CycleGAN can obtain 78.60% accuracy while our CycleGAN-4MBD method achieves the highest 84.28% as a reason of generating high quality images with diversity. We also provide some visual translation comparison results between CycleGAN and our proposed method in Fig. 15. As shown in this figure, our method can handle the imbalanced image translation reasonably. Compared with the original CycleGAN method, the proposed method can generate more realistic image outputs, which leads to a gain of the image classification performance.

Table 6: The classification accuracy of the image classifier model using one dominant category and 5 rare categories.

Method	Data split		Classification accuracy				
	train	test	Vanilla	Flip, resize and crop	Focal loss	CycleGAN	CycleGAN -4MBD
Passion	200	51	98.04%(50)	90.20%(46)	88.24%(45)	92.16%(47)	90.20%(46)
Rose	10	51	35.29%(18)	39.22%(20)	64.71%(33)	70.59%(36)	80.39%(41)
Water Lily	10	51	41.18%(21)	45.10%(23)	62.75%(32)	72.55%(37)	82.35%(42)
Windflower	10	44	38.64%(17)	40.91%(18)	63.64%(28)	72.73%(32)	81.82%(36)
English marigold	10	51	47.06%(24)	52.94%(27)	70.59%(36)	82.83%(42)	86.27%(44)
Tree poppy	10	51	45.10%(23)	47.06%(24)	68.63%(35)	80.39%(41)	84.31%(43)
Average	-	-	51.17%	52.84%	70.23%	78.60%	**84.28**%

4.5.3. The boost for general imbalanced learning tasks

In the real world, the imbalanced settings widely exist: such as the disease classification, fraud detection and so on. The minority class usually contains valuable information. Take the disease classification as an example, the healthy identities are always much more than the abnormal identities, so we pay more attention to the detection of the negative samples. Although it is not an exact cross-species example in medical image processing, there exists many similar features between the healthy and diseased examples. To explore the potential of our method on general imbalanced learning problems, we adopt our method to achieve data augmentation by translating the healthy samples to the diseased samples. We perform the experiments on the ODIR5K dataset[1]. We choose the normal sample as the dominant category and the cataract as the rare category. We perform experiments based on the photographs of the left eye. There are 1580 normal samples and 159 cataract samples, which are labeled. To evaluate the effectiveness of pro-

Fig. 15. The visual translation results of proposed method at imbalanced setting, the left column input represent the input passion flower images. The five columns at the right of the black dotted line show the translated results to the five rare categories. During the translation procedure, our method can achieve the position and number matching after the translation.

posed method, we perform the translating between the two categories. Following the similar setting of Sec.4.5.2, the data split and the classification accuracy are shown in Table.7. We also exhibit the visual translation results between the two categories in Fig.16. By introducing the translated images for the imbalanced setting, we can boost the classification accuracy and improve the ability to recognize the negative samples of rare category.

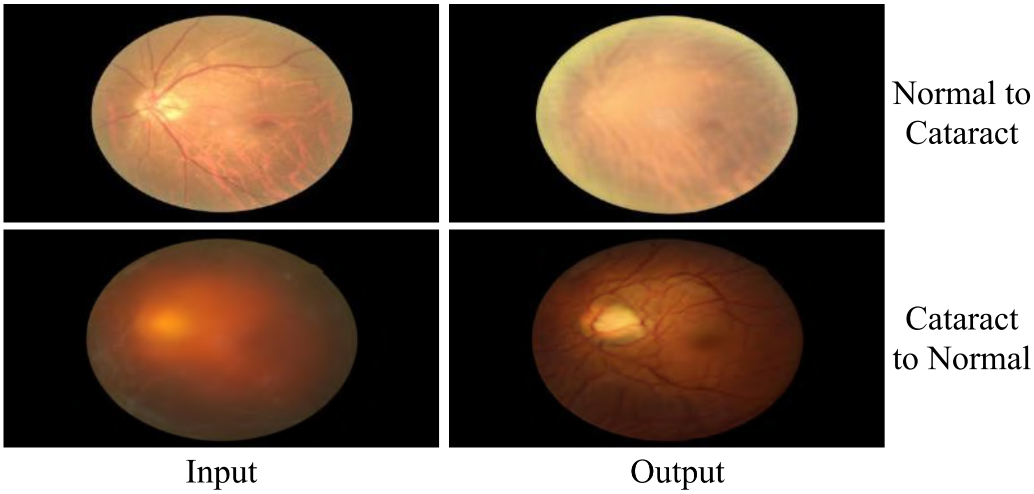


Fig. 16. The visual translation results between the Normal and Cataract fundus photographs of proposed method at imbalanced setting.

Table 7: The classification accuracy of the image classifier model at both two cases: with and without translation. With the translated images from the dominant category, the classification accuracy has been improved a lot.

Method	Data split (normal/cataract)		Classification accuracy (normal/cataract)		
	Train split	Test split	Cross entropy	Focal loss[39]	Cross entropy with translation
Normal and cataract	1480/59	100/100	87.5%(100/75)	91.5%(99/84)	95.0%(100/90)

5. Conclusion

We develop a novel, simple yet effective and efficient multi-branch discriminator (MBD) structure for GANs, leading to high-quality *cross-species* image-to-image translation on the semantic level. We first show the lower bound of MBD and explain the optimal condition of MBD by mathematical analysis. Secondly, our comprehensive experiments show that the proposed MBD structure can effectively improve popular GANs by enhancing the generative ability while efficiently accelerating convergence and reducing parameters dramatically. Finally, we successfully apply the proposed cross-species image-to-image translation techniques on data augmentation tasks and show the potential in the field of imbalanced image recognition.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant Number 61701463; the Key Technology Research and Development Program of Shandong (Public welfare) under Grant Number 2019GHY112041. Ziqiang Zheng and Zhibin Yu contributed equally to this article.

References

[1] , . Ocular disease intelligent recognition odir-5k URL: `https://odir2019.grand-challenge.org/`.

[2] Albuquerque, I., Monteiro, J., Doan, T., Considine, B., Falk, T., Mitliagkas, I., 2019. Multi-objective training of generative adversarial networks with multiple discriminators. arXiv preprint arXiv:1901.08680 .

[3] Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks, in: ICML, pp. 214–223.

[4] Aurelio, Y.S., de Almeida, G.M., de Castro, C.L., Braga, A.P., 2019. Learning from imbalanced data sets with weighted cross-entropy function. Neural Processing Letters 50, 1937–1949.

[5] Bau, D., Zhu, J.Y., Strobelt, H., Zhou, B., Tenenbaum, J.B., Freeman, W.T., Torralba, A., 2018. Gan dissection: Visualizing and understanding generative adversarial networks. arXiv preprint arXiv:1811.10597 .

[6] Benaim, S., Galanti, T., Wolf, L., 2018. Estimating the success of unsupervised image to image translation, in: ECCV, pp. 218–233.

[7] Benaim, S., Wolf, L., 2018. One-shot unsupervised cross domain translation, in: NeurIPS, pp. 2108–2118.

[8] Borji, A., 2019. Pros and cons of GAN evaluation measures. CVIU 179, 41–65.

[9] Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C., 2009. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem, in: Pacific-Asia conference on knowledge discovery and data mining, Springer. pp. 475–482.

[10] Chan, C., Ginosar, S., Zhou, T., Efros, A.A., 2018. Everybody dance now. arXiv preprint arXiv:1808.07371 .

[11] Chawla, N.V., Japkowicz, N., Kotcz, A., 2004. Special issue on learning from imbalanced data sets. ACM SIGKDD explorations newsletter 6, 1–6.

[12] Chen, Z., Tong, Y., 2017. Face super-resolution through Wasserstein GANs. arXiv preprint arXiv:1705.02438 .

[13] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J., 2018. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation, in: CVPR, pp. 8789–8797.

[14] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A., 2018. Generative adversarial networks: An overview. IEEE SPM 35, 53–65.

[15] Doan, T., Monteiro, J., Albuquerque, I., Mazoure, B., Durand, A., Pineau, J., Hjelm, R.D., 2018. Online adaptative curriculum learning for GANs. arXiv preprint arXiv:1808.00020 .

[16] Durugkar, I., Gemp, I., Mahadevan, S., 2017. Generative multi-adversarial networks, in: ICLR, pp. 1–14.

[17] Elson, J., Douceur, J.R., Howell, J., Saul, J., 2007. Asirra: a CAPTCHA that exploits interest-aligned manual image categorization, in: ACM CCS, pp. 366–374.

[18] Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55, 119–139.

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

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

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

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

[23] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium, in: NIPS, pp. 6626–6637.

[24] Hong, Y., Hwang, U., Yoo, J., Yoon, S., 2019. How generative adversarial networks and their variants work: An overview. ACM CSUR 52, 10:1–10:43.

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

[26] Huang, H., Yu, P.S., Wang, C., 2018a. An introduction to image synthesis with generative adversarial nets. arXiv preprint arXiv:1803.04469 .

[27] Huang, X., Liu, M.Y., Belongie, S., Kautz, J., 2018b. Multimodal unsupervised image-to-image translation, in: ECCV, pp. 172–189.

[28] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: CVPR, pp. 1125–1134.

[29] Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. Journal of Big Data 6, 27.

[30] Kearns, M., Valiant, L., 1994. Cryptographic limitations on learning boolean formulae and finite automata. JACM 41, 67–95.

[31] Khan, S.H., Hayat, M., Bennamoun, M., Sohel, F.A., Togneri, R., 2017. Cost-sensitive learning of deep feature representations from imbalanced data. IEEE transactions on neural networks and learning systems 29, 3573–3587.

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

[33] Kingma, D.P., Welling, M., 2013. Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114 .

[34] Kurach, K., Lucic, M., Zhai, X., Michalski, M., Gelly, S., 2018. The GAN landscape: Losses, architectures, regularization, and normalization. arXiv preprint arXiv:1807.04720 .

[35] Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H., Hua, G., 2016. Labeled faces in the wild: A survey, in: Advances in Face Detection and Facial Image Analysis. Springer, pp. 189–248.

[36] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network, in: CVPR, pp. 4681–4690.

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

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

[39] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

[40] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88.

[41] Liu, Z., Luo, P., Wang, X., Tang, X., 2015. Deep learning face attributes in the wild, in: ICCV, pp. 3730–3738.

[42] Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: CVPR, pp. 3431–3440.

[43] Lu, Y., 2018. Cross domain image generation through latent space exploration with adversarial loss. arXiv preprint arXiv:1805.10130 .

[44] Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O., 2018. Are GANs created equal? a large-scale study, in: NeurIPS, pp. 698–707.

[45] Luo, Y., Xu, Y., Ji, H., 2015. Removing rain from a single image via discriminative sparse coding, in: ICCV, pp. 3397–3405.

[46] Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., 2016. Multi-class generative adversarial networks with the L2 loss function. arXiv preprint arXiv:1611.04076 .

[47] Mason, L., Baxter, J., Bartlett, P.L., Frean, M.R., 2000. Boosting algorithms as gradient descent, in: NIPS, pp. 512–518.

[48] Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 .

[49] Nemoto, K., Hamaguchi, R., Imaizumi, T., Hikosaka, S., 2018. Classification of rare building change using cnn with multi-class focal loss, in: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, IEEE. pp. 4663–4666.

[50] Neyshabur, B., Bhojanapalli, S., Chakrabarti, A., 2017. Stabilizing GAN training with multiple random projections. arXiv preprint arXiv:1705.07831 .

[51] Nilsback, M.E., Zisserman, A., 2008. Automated flower classification over a large number of classes, in: ICVGIP, IEEE. pp. 722–729.

[52] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting, in: CVPR, pp. 2536–2544.

[53] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H., 2016. Generative adversarial text to image synthesis, in: ICML, pp. 1060–1069.

[54] Rezende, D.J., Mohamed, S., Wierstra, D., 2014. Stochastic backpropagation and variational inference in deep latent Gaussian models, in: ICML, pp. 1278–1286.

[55] Schapire, R.E., Freund, Y., 2012. Boosting: Foundations and algorithms. MIT press.

[56] Shmelkov, K., Schmid, C., Alahari, K., 2018. How good is my GAN?, in: ECCV, pp. 213–229.

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

[58] Sønderby, C.K., Caballero, J., Theis, L., Shi, W., Huszár, F., 2017. Amortised map inference for image super-resolution, in: ICLR, pp. 1–17.

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

[60] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C., 2018. A survey on deep transfer learning, in: International conference on artificial neural networks, Springer. pp. 270–279.

[61] Tolstikhin, I.O., Gelly, S., Bousquet, O., Simon-Gabriel, C.J., Schölkopf, B., 2017. AdaGAN: Boosting generative models, in: NIPS, pp. 5424–5433.

[62] Van Hulse, J., Khoshgoftaar, T.M., Napolitano, A., 2007. Experimental perspectives on learning from imbalanced data, in: Proceedings of the 24th international conference on Machine learning, pp. 935–942.

[63] Wang, C., Zheng, H., Yu, Z., Zheng, Z., Gu, Z., Zheng, B., 2018a. Discriminative region proposal adversarial networks for high-quality image-to-image translation, in: ECCV, pp. 770–785.

[64] Wang, H., Cui, Z., Chen, Y., Avidan, M., Abdallah, A.B., Kronzer, A., 2018b. Predicting hospital readmission via cost-sensitive deep learning. IEEE/ACM transactions on computational biology and bioinformatics 15, 1968–1978.

[65] Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., Kennedy, P.J., 2016. Training deep neural networks on imbalanced data sets, in: 2016 international joint conference on neural networks (IJCNN), IEEE. pp. 4368–4374.

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

[67] Wang, T.C., Liu, M.Y., Zhu, J.Y., Yakovenko, N., Tao, A., Kautz, J., Catanzaro, B., 2018d. Video-to-video synthesis, in: NeurIPS, pp. 1152–1164.

[68] Weiss, K., Khoshgoftaar, T.M., Wang, D., 2016. A survey of transfer learning. Journal of Big data 3, 9.

[69] Wilkinson, L., Friendly, M., 2009. The history of the cluster heat map. The American Statistician 63, 179–184.

[70] Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H., 2017. High-resolution image inpainting using multi-scale neural patch synthesis, in: CVPR, pp. 6721–6729.

[71] Yeh, R.A., Chen, C., Yian Lim, T., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N., 2017. Semantic image inpainting with deep generative models, in: CVPR, pp. 5485–5493.

[72] Yi, Z., Zhang, H., Tan, P., Gong, M., 2017. DualGAN: Unsupervised dual learning for image-to-image translation, in: ICCV, pp. 2849–2857.

[73] Zhang, C., Tan, K.C., Ren, R., 2016. Training cost-sensitive deep belief networks on imbalance data problems, in: 2016 international joint conference on neural networks (IJCNN), IEEE. pp. 4362–4367.

[74] Zhang, G., Kan, M., Shan, S., Chen, X., 2018a. Generative adversarial network with spatial attention for face attribute editing, in: ECCV, pp. 417–432.

[75] Zhang, H., Sindagi, V., Patel, V.M., 2017a. Image de-raining using a conditional generative adversarial network. arXiv preprint arXiv:1701.05957 .

[76] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N., 2017b. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks, in: ICCV, pp. 5907–5915.

[77] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N., 2018b. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. IEEE TPAMI .

[78] Zhang, Q., Yang, L.T., Chen, Z., Li, P., 2018c. A survey on deep learning for big data. Information Fusion 42, 146–157.

[79] Zheng, Z., Wang, C., Yu, Z., Zheng, H., Zheng, B., 2018. Instance map based image synthesis with a denoising generative adversarial network. IEEE Access 6, 33654–33665.

[80] Zhou, Z.H., 2012. Ensemble methods: foundations and algorithms. Chapman and Hall/CRC.

[81] Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: ICCV, pp. 2223–2232.

[82] Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E., 2017b. Toward multimodal image-to-image translation, in: NIPS, pp. 465–476.

Appendix

5.1. Implementation details

5.1.1. CycleGAN-4MBD

Our `CycleGAN-4MBD` model consists of two discriminators with 4 branches for each. We adopt the generator network architectures from CycleGAN [81]. The Encoder and the Decoder structures are defined as:

Encoder:

$CI64F7 - CI128F4 - CI256F4 - Res256 - Res256 - Res256 - Res256 - Res256 - Res256$

Decoder:

$-Res256 - Res256 - Res256 - CI128F4 - C64F7 - C3$

Here $CImFn$ means the Convolution-InstanceNorm-ReLU layer with m $n \times n$ spatial filters, and $Res256$ means a residual block with 256 3×3 filters. All residual blocks use instance normalization. The last layer of the decoder uses a Tanh instead of a ReLU as the activation function without instance normalization

to obtain the image generation output. Each branch of the discriminator includes one task: True/False discrimination. The structure of each branch is defined as
Discrimination task:
$C16 - C32 - C64 - C128 - C128 - C1$
We set convolution kernel size 4 and stride 2, all ReLUs in the discriminator are leaky, with slope 0.2. We do not use any activation function at the last layer.

Table 8: Computation evaluation (1K=1000 iterations).

Method	Time (s/1K)	DParams (M)	Species
DRIT	2570.49	27.22	Two
MUNIT	**418.58**	16.54	Two
CycleGAN	450.19	13.92	Two
CycleGAN-2MBD	461.30	6.97	Two
CycleGAN-4MBD	595.91	**3.50**	Two
StarGAN	373.44	45.41	Multi
StarGAN-2MBD	342.13	23.06	Multi
StarGAN-4MBD	**315.42**	**11.89**	Multi

5.1.2. Computation comparison

We finally evaluate all the models for cross-species image translation on computation, and Table 8 lists the results. All models are implemented in the same environment (Intel Xeon E5-2620 v4, 128 GB, 1080 Ti, TensorFlow 1.8.0). It can be seen that DRIT trains and tests both very slow, MUNIT trains faster resorting to joint optimization but the parameter amount for inference is larger, and our CycleGAN-MBD models train a little slower than CycleGAN but test faster with fewer parameters, while StarGAN-MBD models train and test both faster than StarGAN.

5.2. Visualizing ensemble discriminators

For better understanding the working mechanism of different structures of ensemble discriminator, we visualize the feature maps of different CycleGAN-based structures with a cat image (left) and a dog image (right) as input in Fig. 17. It can be seen that both multiple branches and multiple discriminators can learn different sub-tasks for each. However, the `MSD` structures have unclear and repetitive division of labor, the `MD` structures perform better than `MSD` structures with clearer division of labor (*e.g.*, edges and eyes) but still worse than our `MBD` structures, which has more clear and less repetitive labor division. Thus, our `MBD` can tackle high-level (*e.g.*, cross-species) image translation better.

5.3. *Additional experimental comparison*

5.3.1. *Cat⇔Human*

The cat↔human image translation is implemented between Cat2dog and CelebA datasets, Table 9 lists the FID comparison, and Fig. 18 shows the visual translation comparison.

Table 9: FID results on cat↔human image translation.

Method	Cat→Human	Human→Cat
DRIT	161.9762	184.2454
MUNIT	152.7639	108.1680
CycleGAN-4MBD	**139.4375**	**67.8787**

5.3.2. *Dog⇔Flower*

The dog↔flower image translation is implemented between Cat2dog and 102Flowers datasets, Table 10 lists the FID comparison, and Fig. 19 shows the visual translation comparison.

Table 10: FID results on dog↔flower image translation.

Method	Dog→Flower	Flower→Dog
DRIT	145.6682	176.7891
MUNIT	138.1872	86.1801
CycleGAN-4MBD	**118.2336**	**74.2712**

Table 11: FID results on dog↔human image translation.

Method	Dog→Human	Human→Dog
DRIT	**118.8862**	146.2622
MUNIT	142.8348	147.8955
CycleGAN-4MBD	134.0117	**87.1139**

5.3.3. *Dog⇔Human*

The dog↔human image translation is implemented between Cat2dog and CelebA datasets, Table 11 lists the FID comparison, and Fig. 20 shows the visual translation comparison.

5.3.4. More results

We present more results of cat↔dog, cat↔flower (controlled and wild), flower↔human (controlled and wild) in Fig. 21, Fig. 22, Fig. 23, Fig. 24, Fig. 25, respectively.

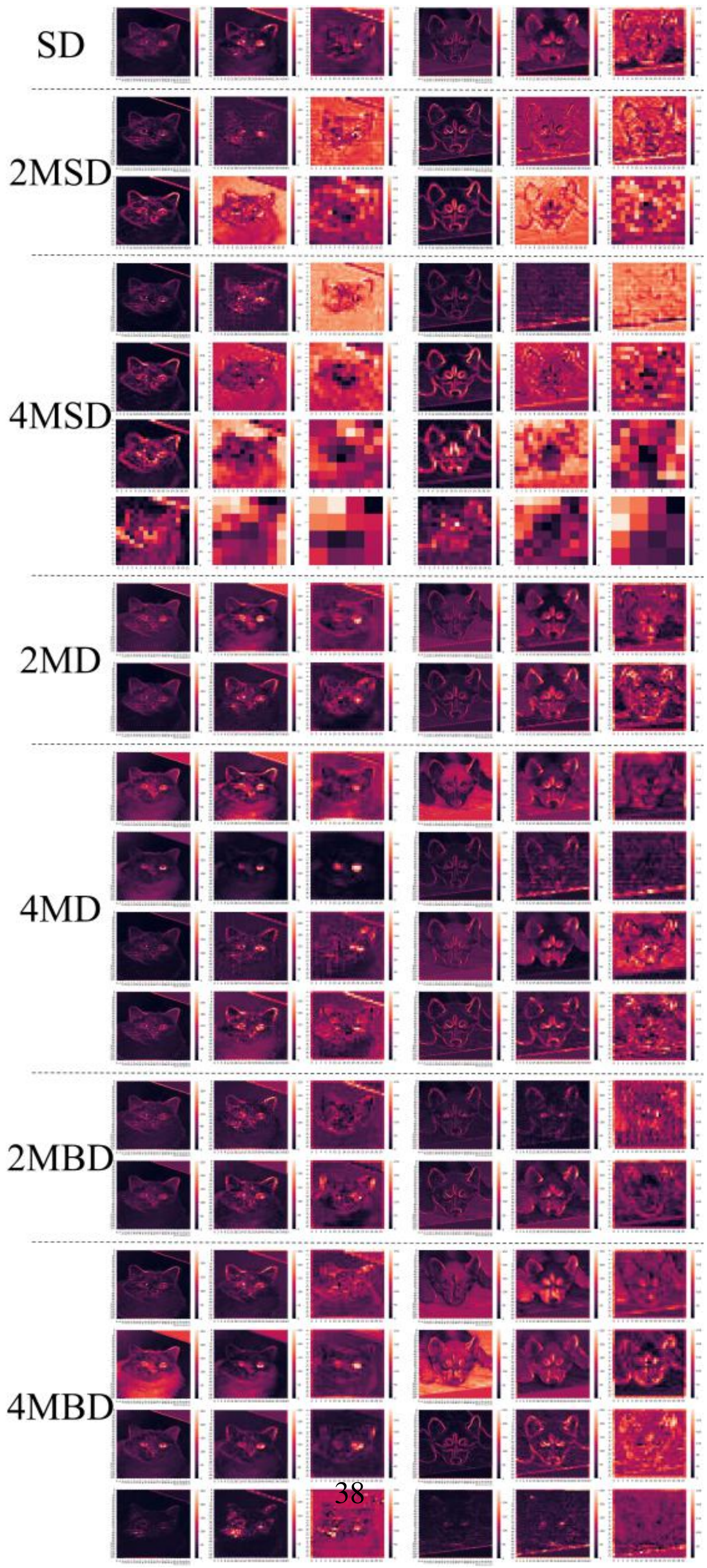


Fig. 17. Visualizing different CycleGAN-based ensemble discriminator structures for comparison.

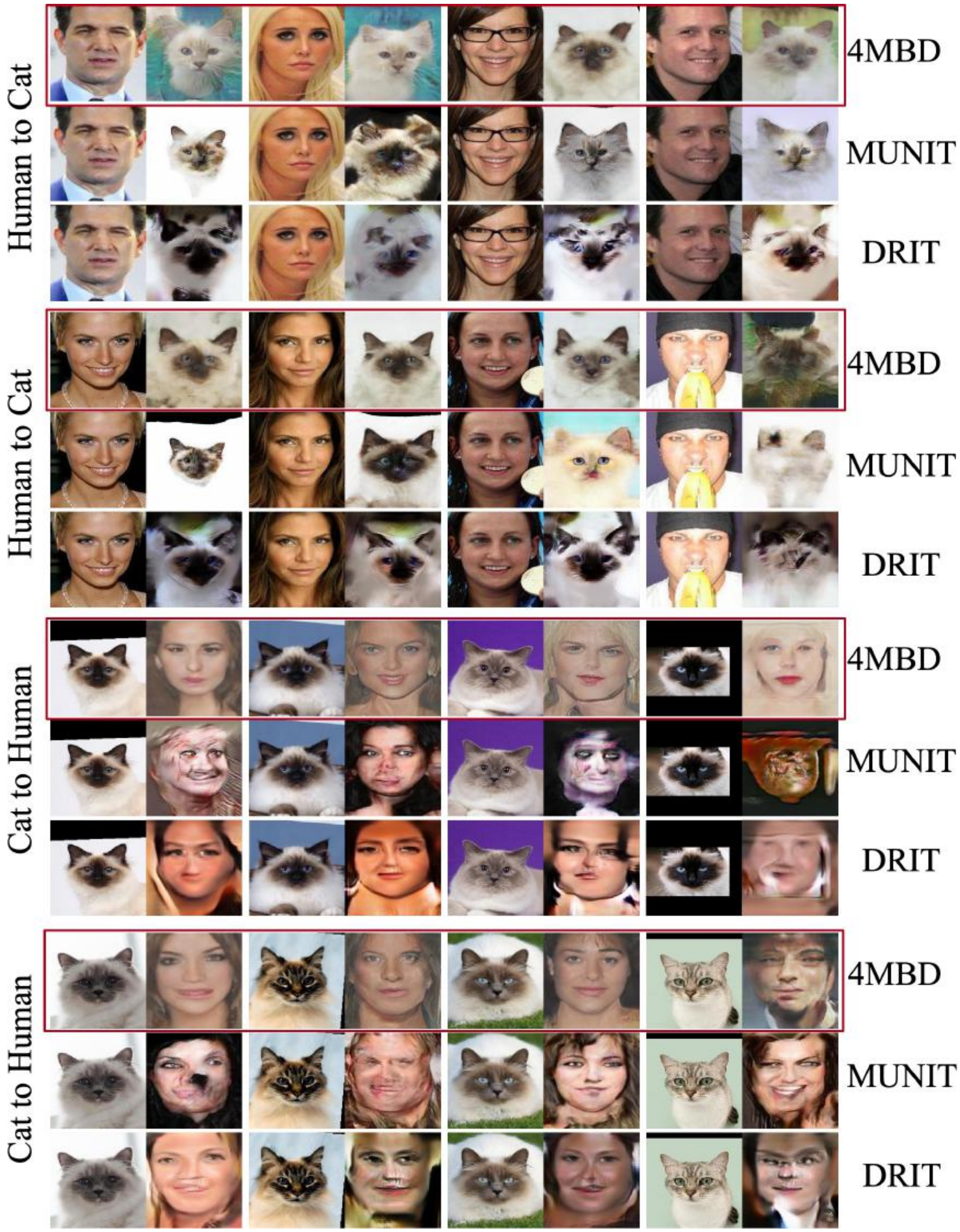


Fig. 18. The cat↔human image translation results of our CycleGAN-4MBD compared to MUNIT and DRIT.

Fig. 19. The dog↔flower image translation results of our CycleGAN-4MBD compared to MUNIT and DRIT.

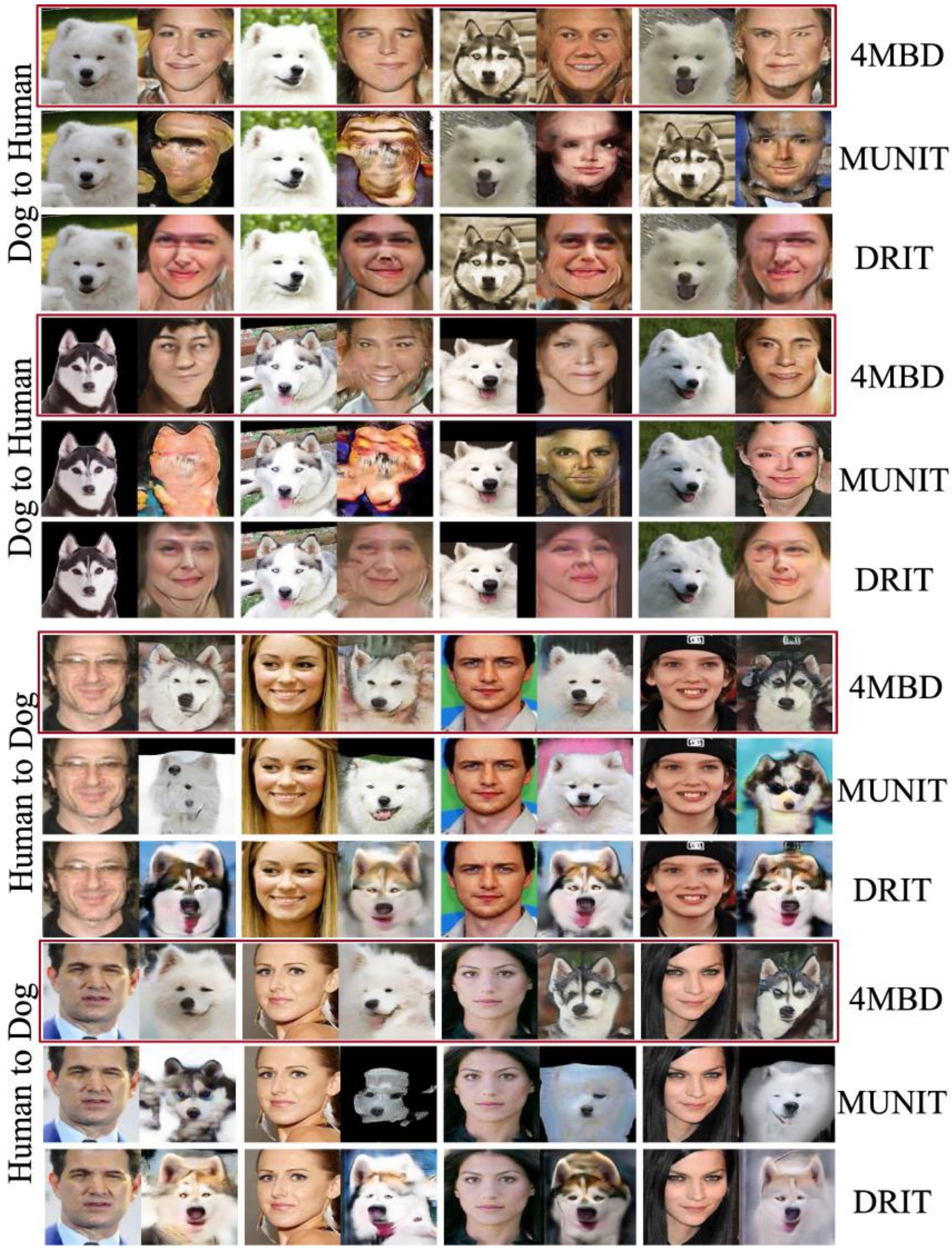


Fig. 20. The dog↔human image translation results of our CycleGAN-4MBD compared to MUNIT and DRIT.

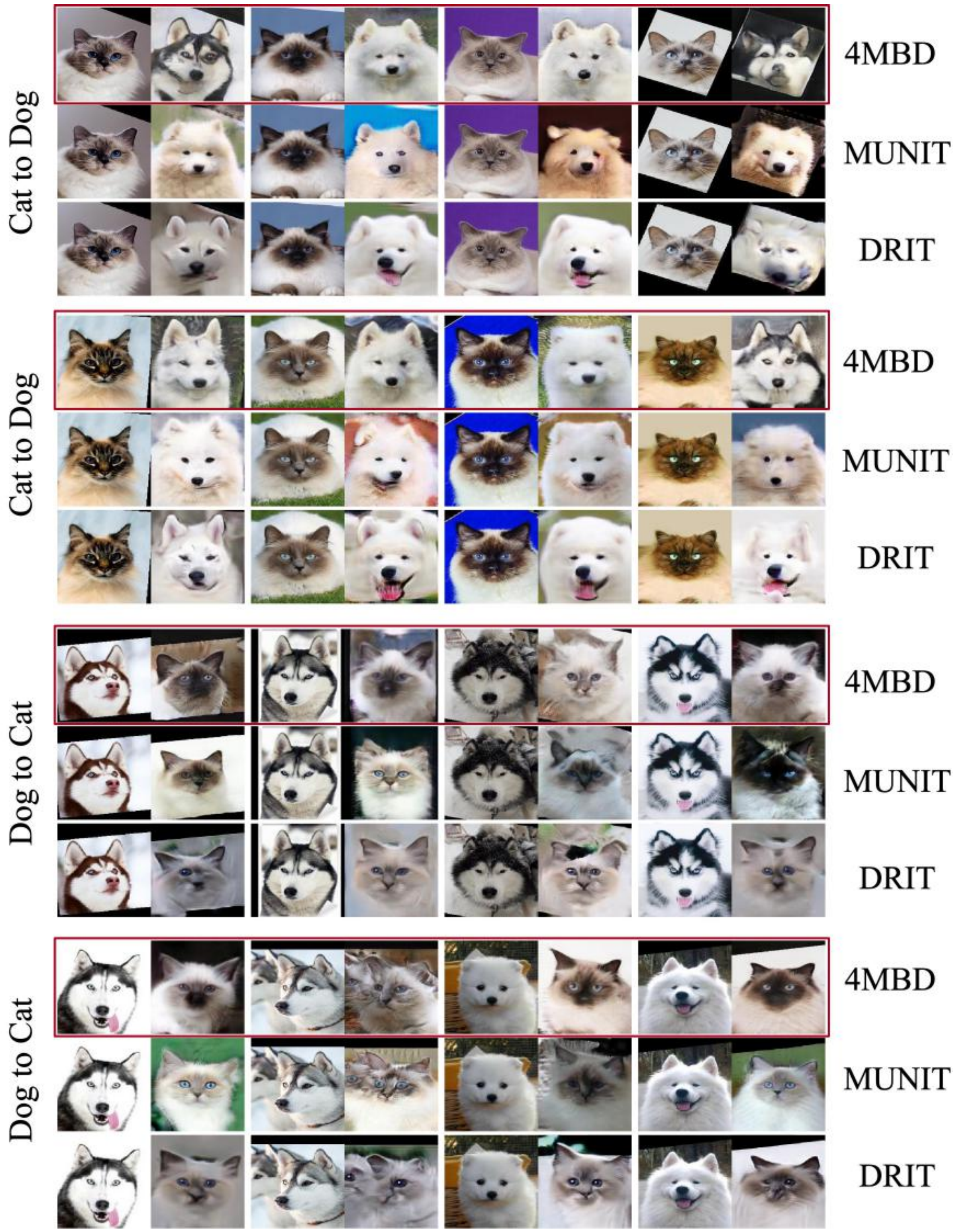


Fig. 21. The cat↔dog image translation results of our CycleGAN-4MBD compared to MUNIT and DRIT.

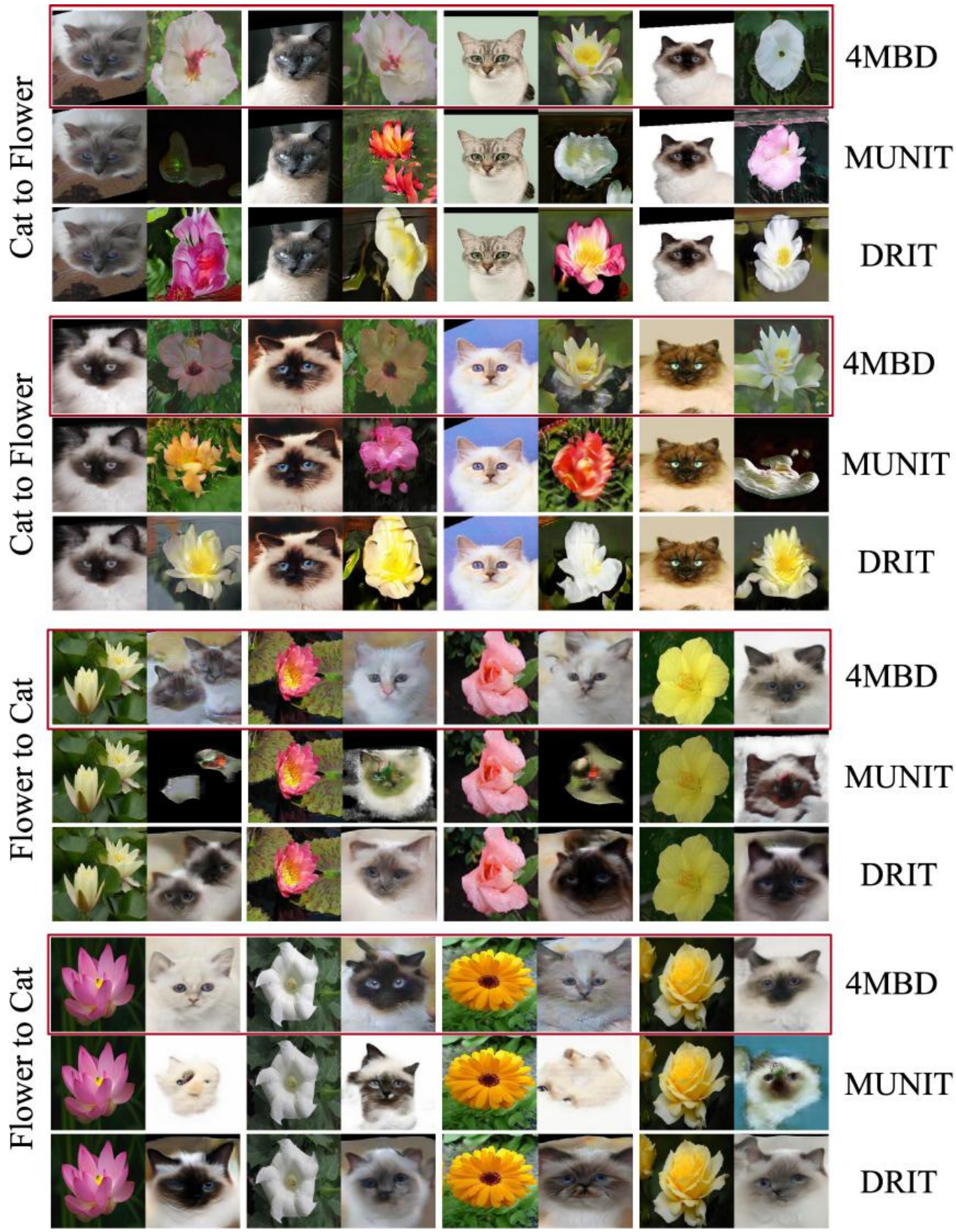


Fig. 22. The cat↔flower image translation results of our CycleGAN-4MBD compared to MUNIT and DRIT.

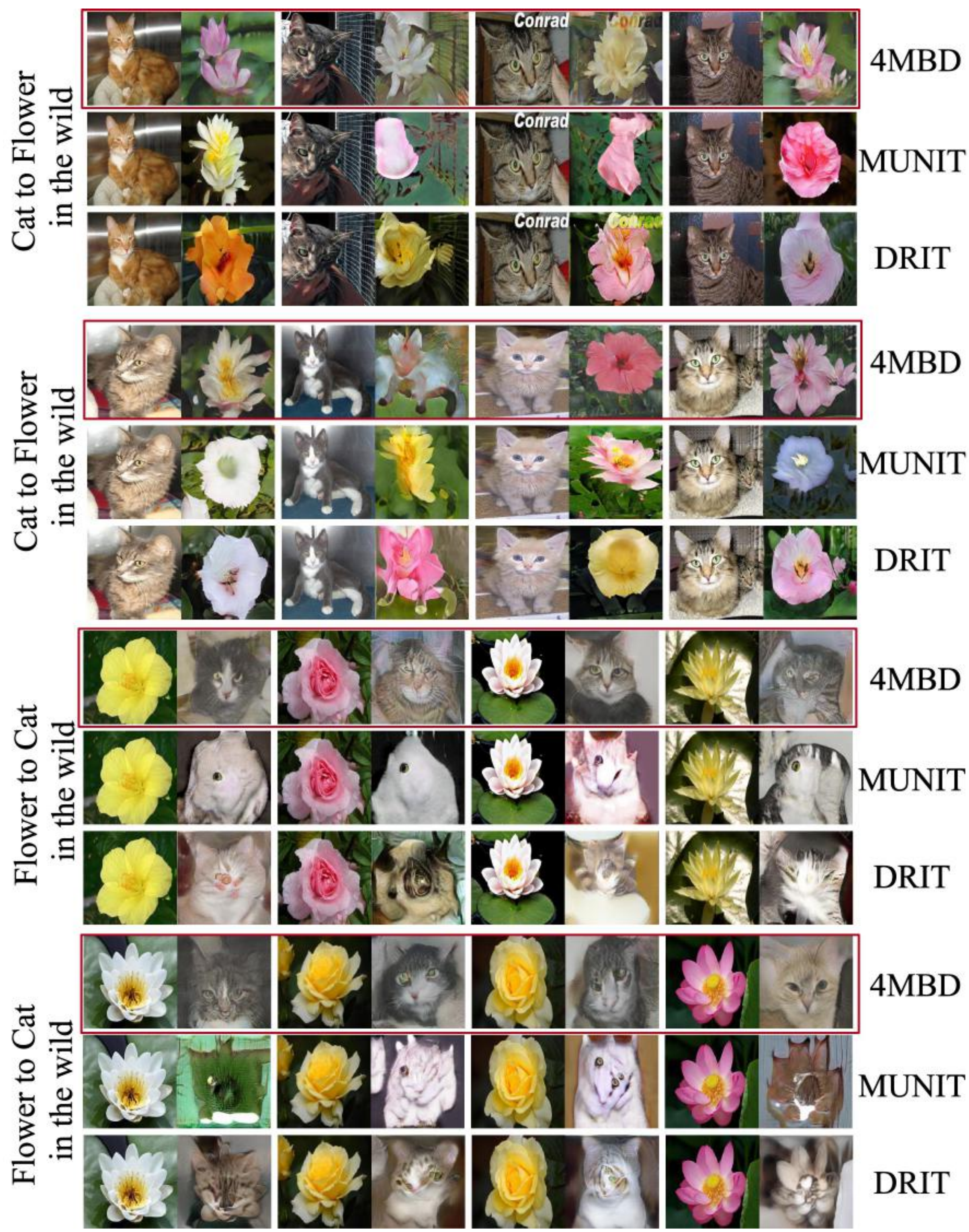


Fig. 23. The wild cat↔flower image translation results of our CycleGAN-4MBD compared to MUNIT and DRIT.

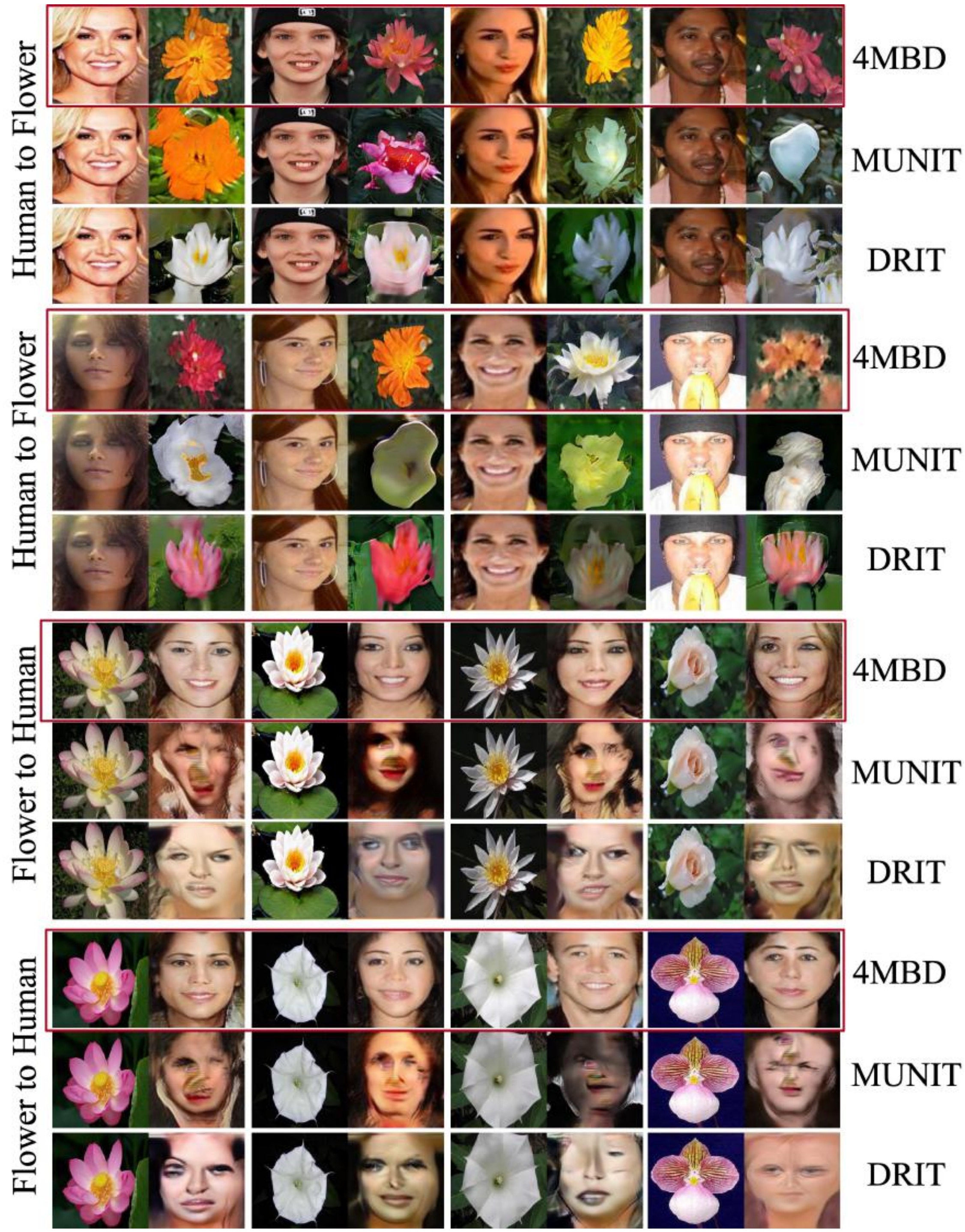


Fig. 24. The flower↔human image translation results of our CycleGAN-4MBD compared to MUNIT and DRIT.

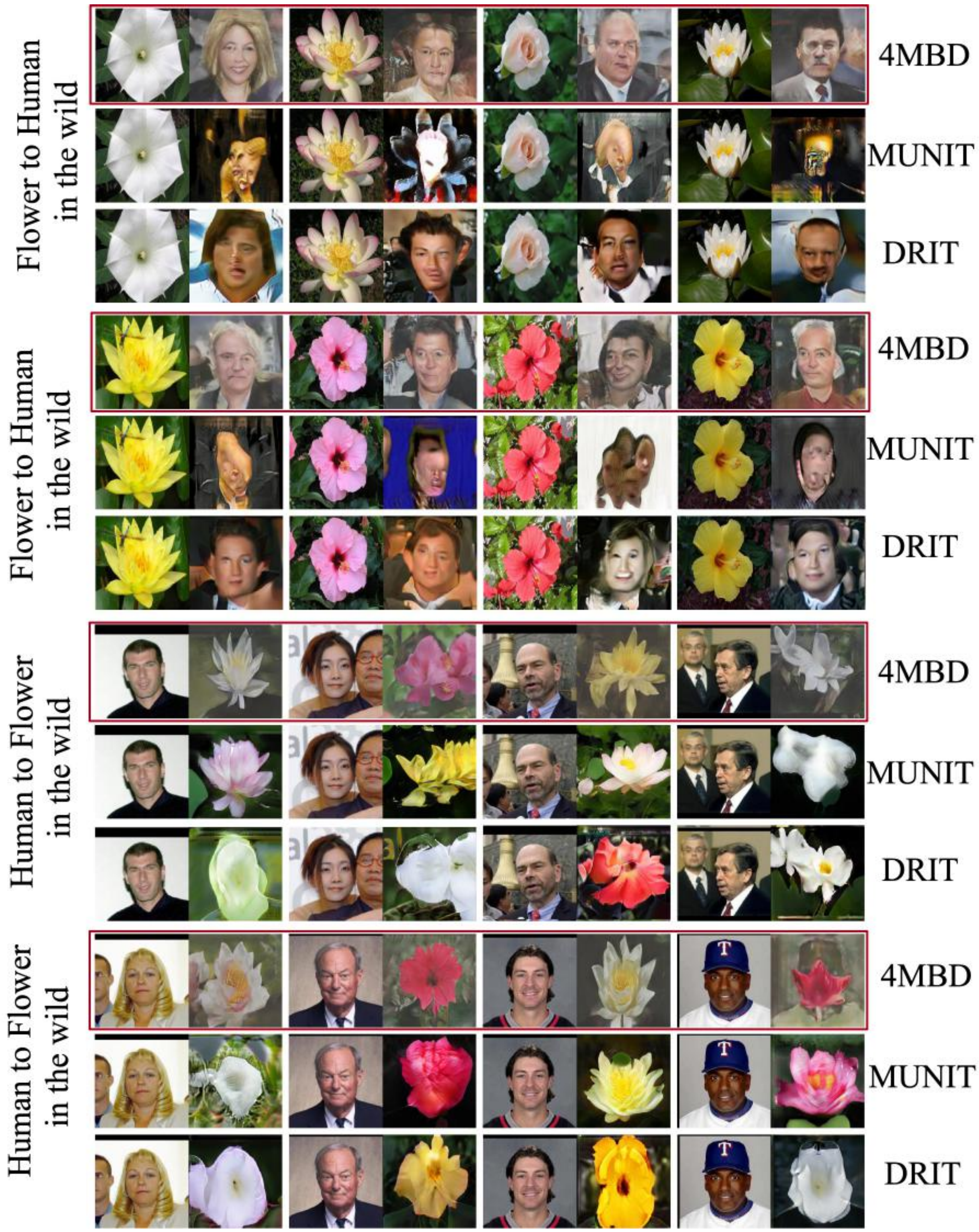


Fig. 25. The wild flower↔human image translation results of our CycleGAN-4MBD compared to MUNIT and DRIT.