IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Instance-Level Image Translation with a Local Discriminator

MINGLE XU[1], JAEHWAN LEE[1], ALVARO FUENTES[1], DONG SUN PARK[2], JUCHENG YANG[3], AND SOOK YOON[4].

[1]Department of Electronics Engineering, Jeonbuk National University, Jeonbuk 54896, Korea
[2]Core Research Institute of Intelligent Robots, Jeonbuk National University, Jeonbuk 54896, Korea
[3]College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin,China
[4]Department of Computer Engineering, Mokpo National University, Jeonnam 58554, Korea

Corresponding author: Dong Sun Park (e-mail: dspark@jbnu.ac.kr) and Sook Yoon (e-mail: syoon@mokpo.ac.kr).

This research is partly supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2019R1A6A1A09031717) and National Research Foundation of Korea (NRF) funded by the Korea government (MSIT) (No. 2020R1A2C2013060). We also thank Zhihui Wang for valuable suggestions.

ABSTRACT Instance-level image translation aims to *only* translate instance of interest and can be operated more finely and flexibly than object-level and holistic-level image translation. However, current algorithms are not suitable to do it since they employ a holistic or object level's discriminator that tends to change the whole image or all instances. To address the issue, we propose a simple yet effective local discriminator, in which the input image is split into two parts, region of interest (ROI) and background. Instance mask is employed to align the ROI and the background is design to be random in a prior distribution to mitigate a divergence between the ROI and the background. In this way, we obtain translated instance with decent margins without artifacts as current algorithms get. Moreover we propose a new architecture to simultaneously realize versatile instance-level image translation. Experimental results prove that our proposed algorithm outperforms the state-of-the-art in position accuracy and background retainment by a clear margin.

INDEX TERMS Image Translation, Local Discriminator, Generative Adversarial Network.

I. INTRODUCTION

A rapid development has been witnessed in **image-to-image (I2I) translation** with generative adversarial networks (GAN) [1]. I2I aims to learn the mapping between two domains. According to the type of dataset, it can be grouped into paired [2] and unpaired [3], [5], [6]. On the other hand, we can factorize I2I into three levels according to the content that we want to translate, holistic-level, object-level and instance-level. In **holistic-level image translation**, holistic image is taken as one domain and expected to be translated, such as style transferring [7] and semantic image synthesis [8], [9] but one natural images commonly include various objects. Fortunately, **object-level image translation** [10], [11], [12], [4], has been proposed to address the issue. It hypothesizes that one image can be split into two parts, interest of object to be translated and background to be retained. In this paper, we extend the assumption that *there are several instances belonging to same domain in one image* and come up with **instance-level image translation**. To be clear, *object* in this paper means specific domain, such as horse or zebra domain, while *instance* denotes each entity for specific object or domain, such as one horse and another horse. For example, object horse include two instances in Figure 1. In horse and zebra case, object level image translation always translates all horses or zebras but instance level image translation allows us only translate one of them or desired horses and zebras.

Differing with object-level image translation, hence, instance-level image translation not only imposes a restriction on the background but also makes use of the difference of individual instances. In the latter one, we can translate part instance(s) of interest and keep other part. In other words, region of interest (ROI) evolves to be individual instance(s) from all instances in specific domain. Meanwhile, it degrades as object-level image translation when our interest are all instances and degrades further as holistic image translation when there is no limitation on the background. Although instance-level image translation is not first proposed in this paper, we emphasize that the meaning of instance-level in this paper is rather distinct from the literature, [15], [13], [14], in which instance-level information is adopted to improve the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3102263, IEEE Access

translation performance for whole image and therefore all of them can be categorized into holistic-level image translation. On the contrary, we assume in this paper that we just translate the chosen instance and hold the other content. We hold that our instance-level image translation would play a role for other applications such as data augmentation, left as our future work. After all, different domains appearing in one image is common and thus, the generated images by instance-level image translation are more close to real images which can be utilized as augmented data to train instance segmentation and object detection.

To achieve the instance-level image translation, one subtle but key issue appears and that is *how to design discriminator's input?* Feeding holistic image to discriminator is obviously problematic as the background will be evaluated and inevitable changed. Currently there are three plausible mechanisms for the above issue, as shown in Figure 1. *(a) Crop and resize ROI*, [13], [14], [15]. The instance is assumed in rectangle shape and hence the algorithm could behave not always reasonably. *(b) Employ two networks to separately extract holistic image and mask features followed by a fusion module*, [16]. It hypothesizes that the networks can be trained to find the spatial relation between the image and mask. Unfortunately, the assumption is not always satisfied. In other words, the background is always looked by the discriminator and the discriminator inevitably push the generator change the background. *(c) Multiply an image with corresponding binary mask in which the background is static and commonly are set zero*. The discriminator is suspected to devote itself on ROI but one question elusively remains. To be clear, the receptive field in the last layer of discriminator can be classified into three kinds, as the colorful bounding boxes in Figure 1 (c). The blue box only covers background, the yellow one only covers instance, but the red one covers background and instance (if the last layer of discriminator is scalar, only the red one exists). Because of the semantic divergence between instance and background, the feature in the red receptive field is not as stable as the blue and the yellow one, which poses a recognition dilemma for discriminator in the margin area. One of the evidences is the artifacts or blur appearing in the edge between instance and background in the synthesized images.

Following the third method as instance spatial information can be utilized accurately than other two, we believe that the translated margin area would become better if the distribution of background is same or similar to the distribution in the ROI. In this case, the dilemma of the three receptive fields mentioned in last paragraph disappears since the fields are subjected to a similar distribution with the background. Empirically, we take the ROI as a random distribution and replace it with the background. As illustrated in Figure 1 (d), we leverage a random background in a prior distribution and fuse it with ROI in spatial-wise. In this way, the divergence between instance and background is diminished and then the recognition dilemma mentioned in last paragraph disappear. Although the random background is very simple, experimental results validate its efficiency.

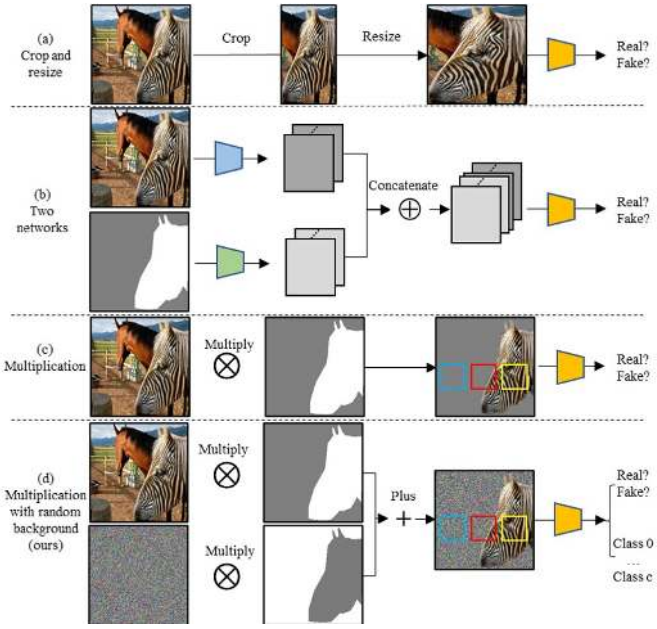

FIGURE 1: The input schemes of local discriminator. The blue, red and yellow bounding boxes represent receptive fields.

Another interesting issue in image translation is versatile generator that one generator is employed to do more than one translation between domains, such as StarGAN [17] in which the emotion of one face can be translated into various emotions. Conversely, non-versatile translation must employ domain-specific generator and discriminator, for example CycleGAN [3]. Obviously, non-versatile image translation requires more networks and parameters and may suffer from overfitting because of imbalance training data. Unfortunately, versatile translation has received much less attraction in object-level and instance-level than holistic-level image translation. One of the main differences is that more inputs are required. Hence, how to design an architecture for generator to absorb three multi-modal inputs, image, mask and target label, is a new challenge. To overcome it, we design a novel encoder to merge label and mask in which we first embed one-hot label into vector and then multiply the vector to the binary mask in a spatial manner, followed by a network to get label and mask feature map. The feature map is then concatenated with image feature map got by another encoder to form a final feature map. With the two encoders, a three multi-modal inputs encoding scheme is achieved and can be directly combined with popular decoder to generate the final translated image.

To summarize, we have following contributions:

1) we propose a very simple yet effective local discriminator taking an extra random background as input, which is compatible with existing algorithms. Armed with the new local discriminator, decent translated instance with competing margin is received;

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3102263, IEEE Access

2) we design a novel architecture for generator to achieve two goals simultaneously, instance-level image translation and versatile generation, to which we design a novel mask and label encoder;
3) to evaluate algorithms of instance-level image translation, we introduce a novel evaluation metric, termed mIoU, which takes position-accuracy into consideration.

II. RELATED WORK

A. IMAGE TRANSLATION

Pix2pix [2] can be regraded as the first model with generative adversarial network (GAN) to do image-to-image translation (I2I) successfully with paired training data and I2I then received a great amount of interests. Although many new algorithms have been proposed, majority of them can be classified as holistic-level image translation, for example, unpaired image translation algorithms, CycleGAN [3], DiscoGAN [5], DualGAN [6], which start from a similar idea that one image will be reconstructed if it undergoes twice translation between two domains. Multi-modal image translation, aiming to get multiple different outputs with same input and received much attentions recently, can also be categoried into holistic-level, such as [18], [19], [20], [21]. Generally, holistic-level image translation hypothesizes that one image just include one domain or one style and tends to change the whole image.

Obviously, the hypothesis is not always satisfied, for example translating object of interest but retaining the background [11]. To address the issue, object-level image translation is designed [11], [12], [10], [22]. Object-level image translation can be factorized into two works, finding the potential object and then do object translation while retaining the background [11]. Therefore, it assumes that all instances in the domain must be translated into another domain. In this paper, we further loose the assumption to make individual instance have its own freedom and then explicitly raise a new challenge, namely instance-level image translation.

In instance-level image translation, it is not compulsory to eliminate all instances in source domain. In fact, source and target domain often appear simultaneously, such as horse and zebra standing together, dog and cat playing with each other. Therefore, instance-level image translation is much closer to images taken by our human and can be operated or controlled in a fine and precise way. Even InstaGAN [16], assuming that using prior information such as instance masks contributes to distinguish different instances, can be utilized to do instance-level translation, we rethink using mask and image in an effective way to have a higher position-accuracy in translation process and realize instance-level and versatile image translation simultaneously.

Instance-level image translation also appears in existing papers. Shen et.al believe that a natural image tends to include multiple objects and thinking them in a same way could incur issue for every instance has its own style and attribute [13]. Meanwhile, instance difference is gradually getting more popular to be employed in image translation, [13], [15], [14]. But the meaning of instance-level or instance on these paper is rather different from ours. In those papers, instance-level is adopted to get a better holistic image translation with distinctive instances but in our case the background is required to keep as possible and be coherent in the margin with new object.

B. GENERATIVE ADVERSARIAL NETWORKS

Vanilla generative adversarial network (GAN) can be adopted on image generation with random noise [1]. Its adapted version, Conditional GANs (CGAN) [23], is widely used to achieve many interesting applications. On the basis of CGAN, we introduce multimodal conditions in our algorithm, RGB image, binary mask and one-hot label. To merge those conditions, two encoders are introduced. Therefore, our algorithm is a natural extension of CGAN yet to be more controllable and practical.

On the other hand, original GAN loss is developed to minimize the distance between two holistic-image distributions such as WGAN [25] and LSGAN [26]. But instance-level image translation essentially requires local cognition and translation rather than checking the whole image. Hence, the input of discriminator should be adaptive accordingly. Unfortunately, this issue is too subtle to be found. Thus, we propose a very simple but effective strategy to design discriminator's input, as shown in Figure. 1 (d).

C. VERSATILE TRANSLATION AND MULTI-TASK LEARNING

It is very common to have multiple domains in image translation. To address it, a straightforward way is employing domain-wise generator and discriminator such as CycleGAN [3]. To compare, a versatile generator is domain-agnostic who asks less parameters and further eases overfitting. From multitask learning's viewpoint [27], one translation for a domain-pair may be related with another pair. Therefore, putting them together can receive a better learning process. ACGAN [29] is one seminal work to achieve multi-domain image generation but it is not for image translation, in which an auxiliary domain classifier is employed. For image translation, starGAN [17] is the first model to realize versatile translation by using label embedding and sharing other modules. In StarGAN, the emotions of one face can be translated into different emotions with assigned input labels. Meanwhile, an auxiliary object-classifier is deployed along with real and fake prediction in the discriminator. Similarly, starGAN v2 [40] and [39] adopted a two-stage discriminator, in which a shared feature extractor is the first stage while in the second stage domain specific real or fake classifiers are deployed.

Following ACGAN [29] and StarGAN [17] et.al, we adopt a discriminator with a simple auxiliary object-classifier, as displayed in Figure. 1 (d). But, to the best of our knowledge, we are the first to achieve instance-level image translation and versatile translation simultaneously.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3102263, IEEE Access

III. PRELIMINARY

As defined before, the instance-level image translation in this paper aims to translate the instance to a specific domain but retain the background, which differs from the usage that instance-level is considered but for holistic image translation, [15], [13], [14].

It is well-known that GAN loss can be deployed to do image translation [1], [26], [30], [31]. Originally, holistic image is taken as one domain and the whole image is expected to be translated into another domain. To achieve it, original GAN loss in (1) can be successfully applied (here we use the Least Squares GAN loss [26] since it shows its priority in training stability). Obviously, as discriminator checks the holistic image the holistic-level GAN loss pushes the generator change the whole image.

$$L_{GAN} = \mathbb{E}_y\{(D(y)-1)^2\} + \mathbb{E}_{y'}\{D(y')^2\}, \quad (1)$$

where y is real image, y' is generated image and D is the discriminator.

To translate a local area, cropping and resizing trick, as shown in Figure 1 (a), was introduced in [13], [14], [15]. The local area given by a rectangle is then hypothesized as a domain and translated into another domain. Rectangle is not good enough to stand for instance since some instances overlap together or background is included. Take the bounding box in Figure 1 (a) as example, two horses are included in one box and the undesired horse would be also translated into zebra.

Another possible way is using two networks to learn the connection between image and mask as displayed in Figure 1 (b), [16], in which the spatial connection is supposed to be learned after training. Unfortunately, experimental results suggest that the connection is not always recognized. To be more specific, some instances that we do not want to change are changed and some instances that we want to translate are not translated.

To learn the spatial connection better, AGGAN [12] firstly employed a new type of GAN loss as showed in Figure 1 (c), multiplying mask with image. If region of interest (ROI) is explicitly given, AGGAN loss can be rewrote as (2), in which the discriminator is expected to just evaluate the ROI. Unfortunately, there are potential deficiencies. To explain clearly, we can cast the receptive field in the last layer of discriminator into three categories: only background, only translated instance and the combination of background and the instance, as the colorful boxes. (If the discriminator is scalar, there is only the last receptive field.) The main issue is from the combination receptive field in that background and the instance have different statistical features. The main feature of the receptive field is based on the percentage of the instance, which results in a learning instability. The instability incurs empirically artifacts or blurring in the margin of the translated instance.

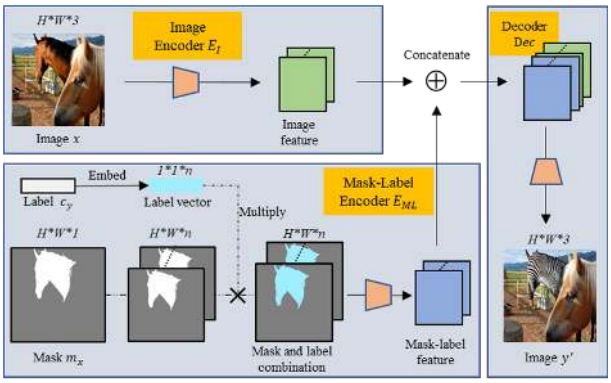

FIGURE 2: Proposed generator's framework from horse to zebra domain. It consists of three sub-models. Image encoder E_I extracts necessary features from the original image, mask-label encoder E_{ML} combines desired label and assigning location denoted by binary mask into features, and decoder Dec generates the output.

$$L_{AGGAN} = \mathbb{E}_y\{(D(y * m_y)-1)^2\} + \mathbb{E}_{y'}\{D(y' * m_{y'})^2\}, \quad (2)$$

in which m_y and $m_{y'}$ are the corresponding mask of y and y'.

To address the issue, we believe that if we can set background similar to instance the instability will disappear. We assume that the pixel value inside the instance is subjected to a random distribution and then sample background from the distribution. As displayed in (3) and illustrated in (d) of Figure 1, we put forward a local GAN loss function termed RBGAN (Random Background GAN). By sampling from a random noise, background of real and generated images are supposed to be a same distribution with the instance. In this way, the receptive filed in the last layer of discriminator is invariant to the location or size of the background, which makes the discriminator and generator easier to be trained.

$$L_{RBGAN} = \mathbb{E}_y\{(D(y * m_y + \mathcal{N} * (1-m_y))-1)^2\} + \mathbb{E}_{y'}\{D(y' * m_{y'} + \mathcal{N} * (1-m_{y'}))^2\}, \quad (3)$$

where $\mathcal{N}$ is a random noise matrix, same size with image y and y'.

IV. PROPOSED ALGORITHM

In this section, we introduce our versatile generator's architecture, training recipes for generator and discriminator, followed by metrics to evaluate instance-level image translation.

A. PROPOSED GENERATOR ARCHITECTURE

To achieve two goals simultaneously, instance-level and versatile image translation, an extra input, mask to show the region of interest (ROI), should be considered except target label and translating image. One of the challenges is how

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3102263, IEEE Access

to merge three information spatially. To address it, a new encoder for fusing target label and the ROI is proposed and explained in the next paragraph.

Figure 2 illustrates our versatile generator's architecture, divided into three parts. The first part image encoder, E_I, encodes input images into feature. Secondly, mask-label encoder E_{ML} merges desired label c_y into specific location denoted by binary mask m_x, in which one means that the pixel belongs to ROI. For example in Figure. 2, only the left horse is expected to be translated and the right horse and other background are required to be kept. In E_{ML}, the label c_y in one-hot format is embedded to a vector in n dimension meanwhile m_x is expanded n times in channel-wise. Then the label vector is multiplied with the expanded mask in every spatial position. In this way, label information is combined with ROI. Followed by several convolution layers, the mask and label combination becomes mask-label feature as the output of E_{ML}. By concatenating the image feature from E_I and the mask-label feature from E_{ML}, a triplet-feature is obtained. Finally, decoder Dec takes the triplet-feature as input and outputs a translated image. Mathematically, (4) is utilized to express the generation process from x to y domain. (To simplify, other types of translation are omitted)

$$y' = Dec\{E_I(x) \oplus E_{ML}(m_x, c_y)\}, \tag{4}$$

where y' is the generated image from the original image x and m_x denotes the binary mask corresponding to x. c_y is the desired label that we want to have in the translated image and an identity translation appears if we just replace c_x with c_y. $\oplus$ denotes feature map's concatenation in channel-wise. Besides, we can imagine that the instance in y' has same mask with x and we will just use $m_{y'}$ for y' in the following notations.

B. LOSS FUNCTION

Except for proposed RBGAN loss mentioned in the last section, an auxiliary object-classifier loss is borrowed to achieve instance-level and versatile image translation. Similar to [29], [17], it shares all computations with discriminator except the last layer and output label prediction c', which merely increases a slight computation burden. Softmax is adopted to compute the classification loss.

$$L_{cls} = softmax(c', c_t), \tag{5}$$

where c' and c_t are the predicted and target class, respectively.

Different with current algorithms with classifier in discriminator, translated image from one domain to another domain is not classified to update the classification network in discriminator since it is far from real domain images, otherwise the classification network would be interrupted in training process. This is illustrated in Figure 3 in which y' is not utilized to do classification. On the contrary, an identity-translated image such as x'' in Figure 3 is extra added to ease the generator as it can converge quickly to its own with the identity loss introduced latter. With this new training scheme, we can obtain a more stable training for classification.

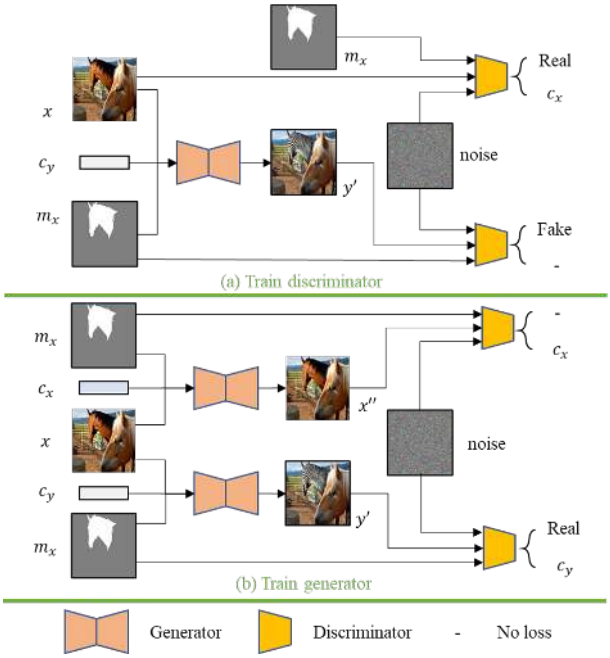

FIGURE 3: Training scheme for generator and discriminator. Real and fake are used as ground truth to compute GAN loss while c_x and c_y are the class ID for the input and the ground truth for discriminator to compute classification loss. When we are not interested in the loss, we use no loss to show the loss will not be adopted to update the generator or discriminator. For example, when training the discriminator with fake image y' as input, we only use GAN loss to update the discriminator while the class c_y denoted as $-$ are not used to update the discriminator.

To retain the background out of the given instance, as shown in (6), L1 norm is applied to compute its difference in pixel-wise between the original image and the translated image. Slightly not same with current algorithms, we do not want to push the background as exactly same as before. We hold that the background should be compatible with the translated instance especially in the margin since no masks is given perfectly.

$$L_{BG} = ||(x - y') * (1 - m_x)||_1. \tag{6}$$

To ease the training of generator, an identity loss, that an instance in target domain should be not changed, is also introduced as (7). Notice that our identity loss also just focuses on the instance other than whole image as other algorithms do in that the background space may not be shared between domains.

$$L_{ident} = ||(y - Dec\{E_I(y) \oplus E_{ML}(m_y, c_y)\}) * m_y||_1. \tag{7}$$

Finally, our full loss function becomes:

$$L_D = L_{RBGAN} + \lambda_{cls} L_{cls}, \tag{8}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3102263, IEEE Access

$$L_G = L_{RBGAN} + \lambda_{cls} L_{cls} + \lambda_{BG} L_{BG} + \lambda_{ident} L_{ident}, \quad (9)$$

where λ_{cls}, λ_{BG} and λ_{ident} are hyper-parameters to balance the losses.

C. EVALUATION METRICS

1) mIoU

To access the fidelity and diversity of the translated images, two of the commonly used assessments are Frechet Inception Distance (FID) [32] and Inception Score (IS) [33], both of them computed in a pre-trained Inception Network [34]. FID and IS aim to compute the feature distribution distance between real images and fake images and intuitively, the smaller distance, the better performance of fake images. Therefor, they are not sensitive to synthetic instance location. Besides, these metrics will lose fairness when target domain exists with source domain in one image. To conclude, FID and IS is not suitable to access instance-level and object-level image translation and new evaluation metrics should be used. Therefore, mask intersection over union (mIoU for simplicity), computed as (10), is proposed in this paper.

$$mIoU = \frac{t_m \cap \sum_{i=1}^{p} g_m^i}{t_m \cup \sum_{i=1}^{p} g_m^i}, \quad (10)$$

where $\cap$ and $\cup$ denote intersection and union of two sets, respectively. t_m denotes the target mask where we want to translate the instance. And g_m means the instance's mask of generated image in the target domain. As tranlation algorithms could not only translate the assigned instance but also other instances even background, g_m may include p instances. To get the generated image's mask g_m, we can annotate the translated images via human labor or use a pre-trained instance segmentation model over the images. To automatically generate translated image's mask, the second method was used in our experiment.

Clearly, mIoU equals one if translation algorithm and the pre-trained model are good enough, which means that the former will translate perfectly for only the given instance and the latter will faultlessly segment the instance's mask. On the other hand, it is close to zero if the given instance is not promisingly translated or the instance in the background is converted. Hence, mIoU is sensitive to translated instance's position as well as the fidelity since the translated instance quality is so low that the pre-trained model could not detect it. A similar idea with our mIoU is the masked classification score in InstaGAN [16]. Though it checks if the expecting position is converted correctly, the score does not check the background.

2) mFID

As discussed before, FID [32] is not suitable to check instance-level image translation. But we find that, with a minor revision, it can be used to evaluate the generated instance. Classic FID computes the feature distribution distance of holistic images which can not distinguish the background and the interested instance. A possible solution for this is that only the interested instance are given and the background are set as zero, in which the discrepancy between the instance and the background leads to unstable impact on the FID. To push the background be similar to the instance, we adopt a random background with instance as input to compute FID, a similar spirit to discriminator's input. Formally, compute the revised FID, termed as mFID, between real image set $\mathcal{R}$ and generated image set $\mathcal{G}$ is as follows:

$$\begin{aligned} mFID(\mathcal{R}, \mathcal{G}) = FID(&\mathcal{R} * m_{\mathcal{R}} + \mathcal{N} * (1 - m_{\mathcal{R}}), \\ &\mathcal{G} * m_{\mathcal{G}} + \mathcal{N} * (1 - m_{\mathcal{G}})), \end{aligned} \quad (11)$$

where $m_{\mathcal{R}}$ and $m_{\mathcal{G}}$ denotes the mask of set $\mathcal{R}$ and set $\mathcal{S}$, respectively.

3) mPSNR and mSSIM

Although mIoU can check the location requirement, it could not explicitly signify the extent to retain the background during instance translation. Hence, we introduce the PSNR and SSIM which have already been widely used as evaluation in image super-resolution and image assessment. Intuitively, PSNR computes the pixel-wise difference between two images while SSIM quantifies the change in structural information (such as local mean, local variance). As we aiming to evaluate the background retain after translating, we use masked ones (only compare the background parts), mPSNR and mSSIM, introduced firstly in [11]. Mathematically, they are computed as Equation. (12) and Equation. (13). In summary, those two metrics are used to assess how the background is retained. The better saving of background, the bigger evaluation values. In the following experiments, we compute the mean values over the all testing samples for those three metrics mentioned.

$$mPSNR = PSNR((1 - m_x) * x, (1 - m_x) * y'). \quad (12)$$

$$mSSIM = SSIM((1 - m_x) * x, (1 - m_x) * y'). \quad (13)$$

V. EXPERIMENTAL RESULTS

A. DATASET

Horse and zebra images were widely employed in many image-to-image papers, [10], [12], [16], but none of them used instance masks. COCO dataset [35] released instance masks for instance segmentation. Therefore, we leveraged COCO dataset to collect horse and zebra images and instance masks. Both image and mask were resized to 256 * 256 in width and height, respectively. We noticed that there were many small instances which could not even be recognized by our human eyes. To evaluate our algorithm and other current methods effectively and fairly, tiny masks had been abandoned when forming dataset. To evaluate instance's size, an index was raised as following: $I_{size} = n_{ins}/(H * W)$. In the equation H and W are the height and width of the resized image, respectively. n_{ins} symbolizes the number of pixels occupied by an instance. The smaller the index, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3102263, IEEE Access

TABLE 1: Number of images and masks for training and testing. n denotes number.

Dataset	Horse	Zebra	Sheep	Cow
n of images in training	1014	797	402	627
n of masks in training	1126	1042	517	748
n of images in testing	253	199	100	156
n of masks in testing	286	265	129	184

smaller the instance. 0.1 was selected as a threshold. In a similar way, sheep and cow are collected. After getting the dataset, we split them into training and testing based on valid mask and Table. 1 shows the dataset.

B. TRAINING DETAILS

To train our algorithm, CycleGAN [3] training recipe was borrowed. Adam optimizer was deployed with a learning rate of 0.0002 in the first 100 epochs and linearly decay learning rate to zero in the second 100 epochs. In order to ease training stability, we adopted a history of generated images to train the discriminator and generator. PatchGAN [36] with 70 * 70 receptive field had also been used.

C. MODEL ANALYSIS

To realize the instance-level image translation, the input of discriminator plays a fundamental role. It consists of two parts, how to combine mask and label and how to ease the training from the divergence between instance and background.

For the first challenge, apart for the multiplication we also considered adding the mask into the image, concatenating the mask into the image, and adopting two individual networks as displayed in (b) of figure 1. Everything else was same and the random background was not used to have a fair comparison. Table. 2 shows the results. The results suggest that multiplication gets much better mIoU which means that the spatial relation between mask and image is learned. Simultaneously, it pushes the discriminator devote itself to translate the instance and retain the background, a lower mFID and higher mPSNR and mSSIM.

For the second issue as discussed before, we assume that the instance is subject to a known distribution and then sample the background from the distribution to remedy the divergence between them. Here two types of distribution is considered, normal and uniform, as well as their ranges. We thought about four cases of the range of random distribution, [-1.0, 1.0], [-0.6, 0.6] [-0.2, 0.2] and without random background. In normal distribution, the values were clipped into the range. The result are displayed in Table. 3. First of all, the FID with same background is also compared with our proposed mFID, adopting a random noise in the background to remedy the divergence between instance and background. From the results, mFID is better than FID, which verifies our assumption, that there is a divergence between instance and background and reducing the divergence contributes to corresponding performance.

TABLE 2: Comparison on the method to combine the mask and image for discriminator. Adding and Concat means adding and concatenate the mask to the image, two-net denotes that separately extracts feature from image and mask, multiply means that multiply the mask with the image (for this experiment, the random background is not used to check the plain multiplication's performance). The lower mFID, the better and the bigger mIoU, mPSNR and mSSIM, the better. The bold value is the best one in the row.

		Adding	Concat	Two-net	Multiply
Fake Zebra	mFID	87.1	101.4	78.3	**48.6**
	mIoU	0.559	0.523	0.547	**0.711**
	mPSNR	24.36	24.33	**26.86**	25.14
	mSSIM	0.900	0.905	0.884	**0.908**
Fake Horse	mFID	144.9	125.7	139.0	**116.8**
	mIoU	0.368	0.396	0.389	**0.413**
	mPSNR	22.98	22.31	21.90	**24.09**
	mSSIM	0.888	0.879	0.856	**0.930**

Besides, one main character of the random noise can be derived from the results that the impact of the random background is slightly invariant to the noise type but related to its range and the target domain. The main reason behind is that they play a key role to reduce the divergence between the instance, its distribution space related with the domain, and the background, highly related with the noise range. In the dataset zebras are observed to have stripes in white and black, a higher distributional range, while most of horses are in bay, a lower distributional range. The experimental results validate that the performance is better when the background is near to the distribution, such as [-1, 1] is better for zebra but [-0.2, 0.2] is better for horse.

To conclude, the results validate our assumption about the divergence between instance and background and sampling the background from a prior random distribution is useful to improve the translation performance. Uniform distribution in absolute 1 and 0.2 space for zebra and horse are adopted in the latter experiments.

D. ABLATION STUDY

Except for the random background, our training scheme is also different to other algorithms. We took our algorithm as baseline and add or reduce schemes to form comparison, as displayed in table 4. In our algorithm, generated image is not employed to update discriminator's ability on classification as [17] and [29] did in that the generated images are from the real images which will make the discriminator confused. In our setting, our classifier can be trained with a high certainty. Otherwise, all performances reduced shown in (d) mainly because the classifier is interrupted by the fake images.

Besides, an identity translation is used in our algorithm to push the generator produce image in correct class. Since an identity loss in (7) is used, identity translation is quick to converge and can guide the generator produce image in correct class. Because of the dependency, we could not use

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3102263, IEEE Access

TABLE 3: The impact of random background's type and range. $\mathcal{U}$ and $\mathcal{N}$ denote uniform and norm distribution, respectively. In the experiments, we only change the variation but not mean for normal distribution since the input image is underwent a normalization with mean as zero. The number * means the range: [-*, *] and the value is truncated for norm distribution. For example, $\mathcal{U}(0.6)$ means an uniform distribution from -0.6 to 0.6. No noise symbolizes without random background. The bold value is the best one in the row of same distribution.

		No noise	$\mathcal{U}(1.0)$	$\mathcal{U}(0.6)$	$\mathcal{U}(0.2)$	$\mathcal{N}(1.0)$	$\mathcal{N}(0.6)$	$\mathcal{N}(0.2)$
	FID	55.5	**36.1**	40.4	53.1	**37.3**	**37.3**	49.6
Fake	mFID	48.6	**32.6**	35.73	49.3	**33.3**	36.1	38.5
Zebra	mIoU	0.711	**0.781**	0.766	0.724	**0.775**	0.766	0.716
	mPSNR	25.14	**25.63**	25.51	25.27	**25.72**	25.44	24.98
	mSSIM	0.908	**0.917**	0.914	0.911	**0.923**	0.919	0.914
	FID	142.2	157.8	141.3	**139.0**	156.1	152.3	**149.0**
Fake	mFID	116.8	126.5	115.6	**113.0**	131.3	**126.4**	130.6
Horse	mIoU	0.413	0.441	0.475	**0.525**	0.439	**0.476**	0.389
	mPSNR	24.09	**24.52**	24.40	24.12	**24.49**	24.24	23.98
	mSSIM	0.930	**0.937**	0.935	0.928	**0.938**	0.933	0.935

TABLE 4: Ablation study of training recipe.

		(a)	(b)	(c)	(d)	(e)
	mFID	**31.8**	35.8	33.9	35.5	53.6
Fake	mIoU	**0.781**	0.760	0.757	0.759	0.698
Zebra	mPSNR	25.92	25.47	24.78	**26.02**	25.88
	mSSIM	**0.926**	0.915	0.899	0.913	0.925
	mFID	**119.8**	120.1	126.2	126.2	132.4
Fake	mIoU	**0.552**	0.468	0.484	0.478	0.438
Horse	mPSNR	24.59	24.32	23.52	24.93	**24.62**
	mSSIM	0.939	0.929	0.907	0.943	**0.945**

(a): baseline, our proposed algorithm.
(b): (a) - identity-translation classification for updating generator.
(c): (b) - identity translation loss.
(d): (a) + generated images to updating classification network.
(e): (a) + cycle-consistency loss [3] to update the generator.

the identity translation alone. As shown in (c), eliminating both of them undermines the generator in both instance and background area. And the identity classification loss contributes to better translated instance, suggested by the fourth column.

Finally, cycle-consistency is proved to be harmful to translate the instance as it gets smaller mFID and mIoU. One of the possible reason could be that the cycle-consistency introduces a too strong restriction between input and generated image as suggested in [37] and [31].

E. COMPARISON WITH STATE-OF-THE-ART

1) Quantitative Comparison

We compare our algorithm to the following algorithms by evaluating the translated instance, mIoU and mFID, and the background, mPSNR and mSSIM. We compare the algorithms in the four domains, zebra and horse, sheep and cow.

CycleGAN [3], utilizes a cycle-consistency loss in pixel-wise and is one of most successful unpaired image translation algorithms.

CLGAN [37], adopts patch contrastive learning, based on a patch instead of whole image.

U-gat-it [38], employs attention mechanism to let the discriminator and generator know where the object is. The above three algorithm belong to object-level image translation.

InstaGAN [16], explicitly uses binary instance mask during image translation but is expected to learn the spatial coherence between instance and mask (two-net showed before).

AGGAN* [12]. Original AGGAN introduces an attention module, an implicit way, to detect which part is background and which part we want to convert. To give a fair comparison, we replaced an explicit mask with the original attention module which was denoted as AGGAN*, in which the background of original image was merged with the translated ROI content. Simultaneously, the discriminator's input was also updated into masked images as shown in (2).

Table 5 shows the results. As can be seen, for both mPSNR and mSSIM, our method outperforms others, which indicates that the proposed model can retain the desired background with the background loss. Although InstaGAN also adopts a similar background loss, the discriminator also push the generator change the background because two networks are trained to learn the spatial connect between mask and image but it is hard to learn, which results in lower mIoU and mPSNR, mSSIM. In terms of mFID, our algorithm achieves a competitive values, which means that it obtains desired target instance. We observed that different target domain requires distinct translation ability, such as all the algorithms get better performance in zebra than horse. Finally, our algorithm obtains much better mIoU, which suggests that our algorithm learns the spatial connection between mask and image and translate the given instance while keep other background. All algorithms fail to be reasonable in cow domain. We suspect that one of the reasons is that this domain requires that the original instance change its shape, which is beyond of the algorithms. Another reason is that the detector is not good enough or its learned patterns are far away from the generator generated.

2) Qualitative Comparison

Translated images of horse $\leftrightarrow$ zebra are shown in figure 4 and sheep $\leftrightarrow$ cow are shown in figure 5. We observed that CycleGAN, CL-GAN, U-gat-it always change the holistic input, such as all the horses are changed into zebras in the third row, which means that current algorithms are not competent to do instance-level image translation and supports the quantitative results in table 5. In InstaGAN, the spatial connection between mask and image is hard to learn and train and tend to change background, as illustrated in the last four rows from zebra to horse. Although AGGAN* can focus on the assigned instance, artifacts and blurring appear in the margin area, such as the generated sheep in the last row. Because of using random noise in the background to reduce the divergence, our algorithm can generate decent target instance in the margin area.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3102263, IEEE Access

(a) x (b) m (c) [3] (d) [37] (e) [38] (f) [16] (g) [12]* (h) Ours

FIGURE 4: Qualitative comparison on horse and zebra domains. [3] is CycleGAN, [37] is CLGAN, [38] is U-gat-it, and [16] is InstaGAN, [12] is AGGAN. Please zoom in to see the detail.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3102263, IEEE Access

(a) x (b) m (c) [3] (d) [37] (e) [38] (f) [16] (g) [12]* (h) Ours

FIGURE 5: Qualitative comparison on sheep and cow domains. [3] is CycleGAN, [37] is CLGAN, [38] is U-gat-it, and [16] is InstaGAN, [12] is AGGAN. Please zoom in to see the detail.

TABLE 5: Quantitative comparison to the state-of-the-art. [3]is CycleGAN, [37] is CLGAN, [38] is U-gat-it, and [16] is InstaGAN, [12] is AGGAN. The bold value is the best one in the row. - means not computation since AGGAN* uses the background of the input image.

		[3]	[37]	[38]	[16]	[12]*	Ours
Fake Zebra	mFID	60.5	104.7	104.0	87.9	45.7	**31.8**
	mIoU	0.524	0.452	0.428	0.613	0.735	**0.781**
	mPSNR	20.62	15.77	18.18	20.85	-	**28.2**
	mSSIM	0.844	0.666	0.742	0.871	-	**0.932**
Fake Horse	mFID	112.5	147.2	127.6	120.2	149.6	**111.0**
	mIoU	0.448	0.331	0.315	0.484	0.450	**0.554**
	mPSNR	20.52	17.01	16.95	19.90	-	**27.36**
	mSSIM	0.810	0.735	0.694	0.834	-	**0.946**
Fake Cow	mFID	155.4	156.6	167.8	157.6	207.0	**153.8**
	mIoU	**0.006**	0.001	0.003	0.001	0.005	0.004
	mPSNR	24.40	18.82	20.47	20.46	-	**26.54**
	mSSIM	0.807	0.778	0.740	0.900	-	**0.907**
Fake Sheep	mFID	**141.6**	152.7	159.8	148.2	160.6	143.7
	mIoU	0.378	0.340	0.293	0.270	0.189	**0.475**
	mPSNR	22.79	18.69	19.62	21.62	-	**27.81**
	mSSIM	0.811	0.808	0.789	0.893	-	**0.926**

VI. CONCLUSION AND FUTURE WORK

To achieve instance-level image translation which requires to translate the given instance and retain the background, we proposed a local discriminator and a versatile generator in this paper. And a novel local discriminator with a random background as input was proposed to mitigate the divergence between the instance and the background. It is validated to have decent margin area in translated instance. Same idea is also useful to evaluate generated images, such as FID. Simultaneously, an mask and label encoder was not trivially designed to achieve instance-level and versatile image translation. Besides, mIoU was proposed that takes the position of translated instance into consideration to evaluate fairly instance-level image translation. Although the shape change as InstaGAN did and diversity of translation images were not considered, our algorithm was proved to be much effective to do instance-level image translation on multiple domains. In the experimental results, our algorithm displayed superiority to literature in terms of the translation performance and the background retainment, which makes image translation more controllable and close to natural image.

For future work, we will investigate how to use our algorithm for other applications, such as a data augmentation method for detection or segmentation. For example, imbalance data of specific objects could mitigate the models' generalization for detection and segmentation, in which our algorithm can be regraded as a data augmentation to increase the number of training samplers close to natural image. Specifically, armed with our algorithm, instance of the classes with plenty samples can be translated into the instance of minor classes to increase the training dataset.

REFERENCES

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

[2] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2017, pp. 1125–1134.

[3] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision 515 (ICCV)*, 2017.

[4] Yuan, L., Chen, D., Hu, H, "Unsupervised object-level image-to-image translation using positional attention bi-flow generative network. " in *IEEE Access*, 7, 2019, pp. 30637-30647.

[5] T. Kim, M. Cha, H. Kim, J. K. Lee, J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning,* vol 70, 2017, pp. 1857–1865.

[6] Z. Yi, H. Zhang, P. Tan, M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation, " in *Proceedings of the IEEE international conference on computer vision,* 2017, pp. 2849–2857.

[7] X. Huang, S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization, " in *Proceedings of the IEEE International Conference on Computer Vision,* 2017, pp. 1501–1510.

[8] T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization, " in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2019, pp. 2337–2346.

[9] R. Li, W. Cao, Q. Jiao, S. Wu, H.-S. Wong, "Simplified unsupervised image translation for semantic segmentation adaptation, " in *Pattern Recognition,* 2020, pp. 107343.

[10] C. Yang, T. Kim, R. Wang, H. Peng, C.-C. J. Kuo, "Show, attend, and translate: Unsupervised image translation with self-regularization and attention," in *IEEE Transactions on Image Processing,* vol 28, 2019, pp. 4845–4856.

[11] X. Chen, C. Xu, X. Yang, D. Tao, "Attention-gan for object transfiguration in wild images, " in *Proceedings of the European Conference on Computer Vision (ECCV),* 2018, pp. 164–180.

[12] Y. A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, K. I. Kim, "Unsupervised attention-guided image-to-image translation, " in *Advances in Neural 540 Information Processing Systems,* 2018, pp. 3693–3703.

[13] Z. Shen, M. Huang, J. Shi, X. Xue, T. S. Huang, "Towards instance-level image-to-image translation, " in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2019, pp. 3683–3692.

[14] D. Bhattacharjee, S. Kim, G. Vizier, M. Salzmann, "Dunit: Detection based unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 2020, pp. 4787–4796.

[15] S. Ma, J. Fu, C. Wen Chen, T. Mei, "Da-gan: Instance-level image translation by deep attention generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2018, pp. 5657–5666.

[16] S. Mo, M. Cho, J. Shin, "Instagan: Instance-aware image-to-image translation," in *ICLR*, 2019.

[17] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2018, pp. 8789–8797.

[18] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, E. Shechtman, "Toward multimodal image-to-image translation," in *Advances in neural information processing systems,* 2017, pp. 465–476.

[19] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, M.-H. Yang, "Diverse image-to-image translation via disentangled representations, in *Proceedings of the European conference on computer vision,* 2018, pp. 35-51.

[20] Q. Mao, H.-Y. Lee, H.-Y. Tseng, S. Ma, M.-H. Yang, "Mode seeking generative adversarial networks for diverse image synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2019, pp. 1429–1437.

[21] W. Xu, K. Shawn, G. Wang, "Toward learning a unified many-to-many mapping for diverse image translation" in *Pattern Recognition,* 2019, pp. 570-580.

[22] H. Emami, M. M. Aliabadi, M. Dong, R. Chinnam, "Spa-gan: Spatial attention gan for image-to-image translation," in *IEEE Transactions on Multimedia,* 2020, pp. 391-401.

[23] M. Mirza, S. Osindero, "Conditional generative adversarial nets," in *arXiv preprint arXiv:1411.1784.*

[24] H. Zhang, V. Sindagi, V. M. Patel, "Image de-raining using a conditional generative adversarial network," in *IEEE transactions on circuits and systems for video technology,* 2019 pp. 3943-3956.

[25] M. Arjovsky, S. Chintala, L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning,* vol 70, 2017, pp. 214–223.

[26] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision,* 2017, pp. 2794–2802.

[27] S. Ruder, "An overview of multi-task learning in deep neural networks," in *arXiv preprint arXiv:1706.05098.*

[28] Y. Choi, Y. Uh, J. Yoo, J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8188–8197.

[29] A. Odena, C. Olah, J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *International conference on machine learning,* 2017, pp. 2642–2651.

[30] T. R. Shaham, T. Dekel, 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.

[31] M. Amodio, S. Krishnaswamy, "Travelgan: Image-to-image translation by transformation vector learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2019, pp. 8983–8992.

[32] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in: *Advances in neural information processing systems,* 2017, pp. 6626-6637.

[33] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems,* 2016, pp. 2234–2242.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2016, pp. 2818-2826.

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision,* Springer, 2014, pp. 740–755.

[36] C. Li, M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *European conference on computer vision,* Springer, 2016, pp. 702–716.

[37] Park, T., Efros, A. A., Zhang, R., Zhu, J. Y, "Contrastive learning for unpaired image-to-image translation." in *European Conference on Computer Vision,* Springer, Cham, 2020.

[38] Kim J, Kim M, Kang H, Lee KH. "U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation," in *International Conference on Learning Representations,* 2019.

[39] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, Hyunjung Shim, "Rethinking the Truly Unsupervised Image-to-Image Translation," in *arXiv preprint arXiv:2006.06500,* 2020.

[40] Choi, Yunjey and Uh, Youngjung and Yoo, Jaejun and Ha, Jung-Woo, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 2020.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3102263, IEEE Access

MINGLE XU received his B.S from Jiangxi Agricultural Universiaty in 2015 and M.S from Shanghai University of Engineering Science in 2018. His main research interests include artificial intelligence and machine learning, computer vision and image understanding. Email: xml@jbnu.ac.kr.

JAEHWAN LEE received the B.S. and M.S. degrees in engineering from Jeonbuk National University, Republic of Korea, in 2012, and 2014, respectively. Currently he is pursuing doctor degree in Jeonbuk National University from 2018 in South Korea. His main research interests include image processing, pattern recognition, and machine learning. Email: dlwo6@jbnu.ac.kr.

ALVARO FUENTES received his B.S. degree in Mechatronics Engineering from the Technical University of the North, Ecuador in 2012, and his M.S. and Ph.D. degrees in Electronics Engineering majoring in Artificial Intelligence and Computer Vision from Jeonbuk National University, South Korea, in 2016 and 2019 respectively. He is currently a Postdoctoral Researcher with the Department of Electronics Engineering at Jeonbuk National University in South Korea. His main research interests include machine learning, deep learning, computer vision, and robotics.

DONG SUN PARK is a professor at the Jeonbuk National University, Republic of Korea. He received his BS from Korea University, Republic of Korea in 1979, and MS and PhD degrees from the University of Missouri, United States in 1984 and 1990. He has published many papers in international conferences and journals. He is a member of IEEE Computer Society. His research interests include computer vision and artificial neural network, especially deep learning. E-mail: dspark@jbnu.ac.kr.

JUCHENG YANG is a full professor in College of Artificial of Intelligence, Tianjin University of Science and Technology, Tianjin, P.R. China. He is number of CCF. He received his B.S. degree from South-Central University for Nationalities, China, MS and PhD degrees from Chonbuk National University, Republic of Korea. He has published over 120 papers in related international journals and conferences, such as IEEE Communications Magazine, IEEE Trans. Industrial Informatics, IEEE Trans. on HMS, Expert Systems with Applications and so on. He has served as editor of five books in biometrics (Intech publisher), he is the associate editor of International Journal of Information Security and Applications (JISA),and as reviewer or editor for international journals such as IEEE Transactions on Information Forensics & Security, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Communications Magazine, and as program committee member of many conferences such as ICNC'06-FSKD'06, JCeSBI'10, IMPRESS'11 and CCBR'13. He is the publicity chair of ICMeCG'10-12. He owns 20 patents in biometrics. His research interests include image processing, biometrics, pattern recognition, and neural networks.

SOOK YOON is a Professor with the Department of Computer Engineering at Mokpo National University in South Korea. She received her Ph.D. degree in Electronics Engineering from Jeonbuk National University, South Korea, in 2003. She was a researcher in Electrical Engineering and Computer Sciences at the University of California, Berkeley, USA, until June 2006. And she joined Mokpo National University in September 2006. She was a visiting scholar at Utah Center of Advanced Imaging Research, University of Utah, USA, from 2013 to 2015. Her main research interests include computer vision, object recognition, machine learning, and biometrics. Email: syoon@mokpo.ac.kr.