Asynchronous Generative Adversarial Network for Asymmetric Unpaired Image-to-Image Translation

Ziqiang Zheng, Yi Bin, Xiaoou Lv, Yang Wu, *Member, IEEE*, Yang Yang, *Senior Member, IEEE*, and Heng Tao Shen, *Fellow, IEEE*

***Abstract*—The unpaired image-to-image translation aims to translate input images from one source domain to some desired outputs in a target domain by learning from unpaired training data. Cycle-consistency constraint provides a general principle to estimate and measure forward and backward mapping functions between two domains. In many cases, the information entropy of images from the two domains is not equal, resulting in an information-rich domain and an information-poor domain. However, existing solutions based on cycle-consistency either completely discard the information asymmetry between the two domains (a common choice), which leads to inferior translation performance for the asymmetric unpaired image-to-image translation, or have to rely on special task-specific designs and introduce extra loss components. These elaborative designs especially for the relatively harder translation direction from the information-poor domain to the information-rich domain ("poor-to-rich" translation) require extra labor and are limited to some specific tasks. In this paper, we propose a novel asynchronous generative adversarial network named Async-GAN, which provides a model-agnostic framework for easily turning symmetrical models into powerful asymmetric counterparts that can handle asymmetric unpaired image-to-image translation much better. The key innovation is to iteratively build gradually-improving *intermediate domains* for generating pseudo paired training samples, which provide stronger full supervision for assisting the poor-to-rich translation. Extensive experiments on various asymmetric unpaired translation tasks demonstrate the superiority of the proposal. Furthermore, the proposed training framework could be extended to various Cycle-GAN solutions and achieve a performance gain.**

***Index Terms*—Asymmetric image translation, generative adversarial networks, unpaired image-to-image translation.**

Manuscript received 2 August 2021; revised 8 November 2021 and 28 December 2021; accepted 28 December 2021. Date of publication 25 February 2022; date of current version 13 July 2023. This work was supported in part by the National Natural Science Foundation of China under Grants U20B2063 and 62102070, and in part by the Dongguan Songshan Lake Introduction Program of Leading Innovative and Entrepreneurial Talents. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Chang Xu. *(Ziqiang Zheng and Yi Bin are contributed equally to this work). (Corresponding authors: Yang Wu; Yang Yang.)*

Ziqiang Zheng, Yi Bin, and Yang Yang are with the Center for Future Media, University of Electronic Science and Technology of China, Chengdu, Sichuan 610056, China (e-mail: zhengziqiang1@gmail.com; yi.bin@hotmail.com; dlyyang@gmail.com).

Xiaoou Lv is with the Department of Statistical Science, University College London, WC1E 6BT London, U.K. (e-mail: xiaoou_lu@outlook.com).

Yang Wu is with the Applied Research Center (ARC), Tencent PCG, Shenzhen 518057, China (e-mail: dylanywu@tencent.com).

Heng Tao Shen is with the Center for Future Media, University of Electronic Science and Technology of China, Chengdu, Sichuan 610056, China, and also with the Peng Cheng Laboratory, Shenzhen, Guangdong 518066, China (e-mail: shenhengtao@hotmail.com).

Digital Object Identifier 10.1109/TMM.2022.3147425

I. Introduction

IMAGE-TO-IMAGE translation refers to a series of vision and graphics problems mapping images from a source domain to a target domain, *e.g.*, from a grey-scale image to a color image. It has been attracting tremendous research interests due to its wide applications, such as 3D-reconstruction and visual design. There are mainly two kinds of image translation categories: paired image translation [1]–[3] and unpaired image translation [4]–[12], where the paired training means the training data from the source and the target domains are one-to-one correspondence, and the unpaired training means there are no such paired correspondences. Paired image translation is popular in the early works [2], [3], [13], [14] which showed very impressive results on many specific tasks due to the pixel-wise supervision from the paired samples. However, such paired training data are difficult to get or inaccessible in most computer vision tasks due to the labor-intensive and time-consuming process of raw data acquisition and labeling, and thus the paired image-to-image translation models can rarely be trained on large-scale data and have very limited applications. Thanks to the application of *cycle-consistency constraint*, for which CycleGAN [4] is the most influential seminal work, unpaired image-to-image translation becomes possible and encouraging results have been achieved on many individual computer vision tasks. Since then, the research progress moves forward to a further and flourishing stage, various approaches, extending the cycle-consistency idea, have been proposed to translate images between different domains [9], [15].

Cycle-consistency constraint aims to preserve the content information of input images under the unpaired training, containing two translation functions, usually denoted as $F(\cdot)$ and $G(\cdot)$, between two domains, where F maps one domain to the other and G does the way back. Though neither F nor G can get direct supervision for the forward translation due to the unpaired training data, their combinations, *i.e.* $F(G(\cdot))$ and $G(F(\cdot))$, shall be able to reconstruct an arbitrary image from its translated counterpart, which means the content information loss under the bi-direct transformations should be as least as possible. This type of self-supervision is the key to cycle-consistency loss, which has led to groundbreaking successes in unpaired image-to-image translation. The bi-direct translations between

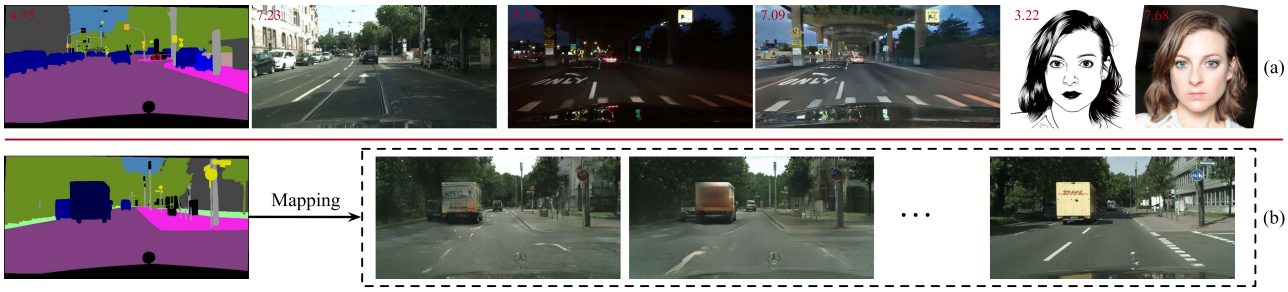

Fig. 1. An intuitive illustration of a) the Shannon entropy of images (the red number in the upper left corner represents the Shannon information entropy); b) multiple photo images share the same semantic instance.

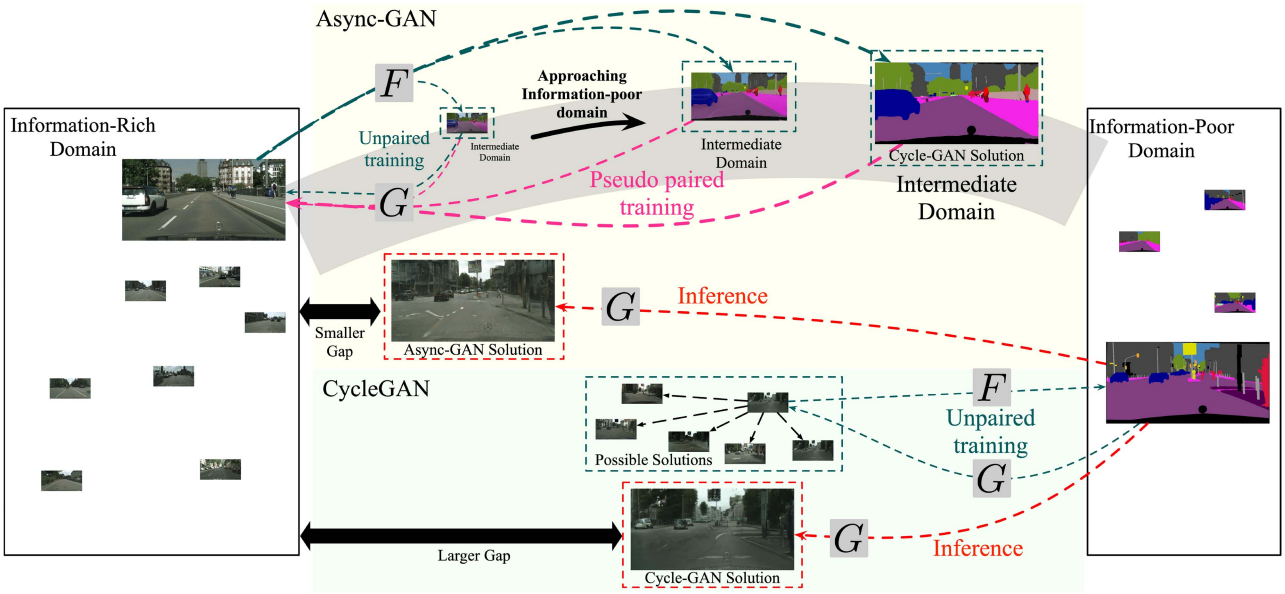

Fig. 2. Illustration of the key ideas of the proposed Asynchronous Generative Adversarial Network (Async-GAN). As shown on the bottom, due to the information asymmetry, there are multiple possible solutions during "poor-to-rich" translation by translator G, and there is usually a large map between its results and the real information-rich domain when training with unpaired data. To make the gap smaller (as shown in the middle), we try to iteratively force the results of "rich-to-poor" translation (via translator F) approach the informative-poor domain by forming *Intermediate Domains*, as shown on the top. Such a strategy naturally provides pseudo paired samples from the intermediate domains and the information-rich domain, which can greatly assist the unpaired training of G by pseudo paired training (the dotted pink lines). By doing so, our Async-GAN can generate significantly better results during inference (shown in the red box) than those from raw cycle-consistency based solutions (shown in the green box).

two domains make it easy to assume F and G are symmetric, which means the difficulties of optimizing F and G are equal. However, practices on various data soon reveal that the results of the two translations are not equally satisfying in many cases though the same amounts of data from the two domains are used. One translation function appears to be significantly much more difficult than the other to be optimized. This phenomenon is due to the information entropy difference between images of the two domains. For example, a color scene image contains much richer information than a quantized semantic segmentation map for the same scene. To make a better understanding, we provide an intuitive comparison in Fig. 1(a) for this information entropy difference, where the Shannon entropy [16], [17] of images is computed. The Shannon entropy is defined as $S = -\sum(p_k * \log(p_k))$, where p_k are frequency/probability of pixels of value k. For the visual images, the higher the entropy, the more information can be transmitted, and the lower the entropy, the less information can be transmitted. And we quantitatively compute the average entropy of images from each visual domain to measure the information complexity. We observe the phenomenon of *information asymmetry* existing in many other image-to-image translation tasks including photo $\leftrightarrow$ portrait [18], photo $\leftrightarrow$ sketch [19], and RGB $\leftrightarrow$ thermal [20] and so on [7], [21]. For simplicity, we call the two domains an information-rich domain and an information-poor domain to represent information asymmetry.

The translation from the information-poor domain to the information-rich domain is generally harder than the inverse translation, as it requires information increase which faces more uncertainty. As shown in Fig. 2(b), one segmentation map may be translated to multiple RGB images, because different RGB images with various local textures and colors could share the same segmentation instance map. Though one may argue that one RGB image may also have multiple reasonable

segmentation maps, the feasible result space of the semantic segmentation maps is much smaller than the RGB images. Such an asymmetry also explains the relative hardness of information-poor domain to information-rich domain translation. In an unpaired image-to-image translation task from the information-poor domain to the information-rich domain ("poor-to-rich"), to make the best out of cycle-consistency, existing solutions generally follow a common strategy of making the whole translation model itself asymmetric by somehow forcing the "poor-to-rich" translation part to be heavier than its "rich-to-poor" counterpart. The *model asymmetry* has been explored in three different directions to enhance the the "poor-to-rich" translation part: increasing its capacity (usually via a more complex structure) [22], adding new constraints [18], [20], [23], and introducing extra explicit conditions [24] or latent variables [25]. Besides being computationally expensive, all these efforts need empirical and task-specific design with extra variables/constraints and hyper-parameters.

In this paper, we introduce a novel strategy that is simple, general, and free of the drawbacks of existing solutions. Instead of working on model asymmetry, our strategy is *to directly make use of information asymmetry*. The key idea is to let the easier and more effective "rich-to-poor" translation guide "poor-to-rich" translation via "pseudo-sample-pair" generation. By doing so, the difficult "poor-to-rich" translation can be optimized with both unpaired data and pseudo paired data, leading to superior results. In greater detail, our proposed novel solution is based on the asynchronous learning of Generative Adversarial Networks, which is called Asynchronous Generative Adversarial Network (or Async-GAN for short). Async-GAN learns a reliable "rich-to-poor" translator first and then utilizes it to generate images that are expected to be in the information-poor domain. Since the weak cycle-consistency loss and adversarial loss cannot ensure that the generated images belong to the information-poor domain or follow the real sample distribution, we treat the actual domain that they stay in as an *intermediate domain*. Thanks to the translation, samples in the information-rich domain and the intermediate domain are naturally paired. Then such generated sample pairs can be treated as pseudo "poor-to-rich" paired data for assisting the training of the "poor-to-rich" translator (denoted by G), which in turn benefits the refinement of the "rich-to-poor" translator (denoted by F) as these two translators are tied by cycle-consistency, resulting in a better intermediate domain. Such an asynchronous learning process can iterate multiple times so that the intermediate domain can gradually approach the information-poor domain. Such key ideas and their advantage are illustrated in Fig. 2. Please note that our asynchronous learning strategy may be incorporated in any existing cycle-consistency based framework to boost its performance. In this work, we just show a few examples (mainly with vanilla CycleGAN [4] and the multi-modal MUNIT [5]) to demonstrate the superiority of Async-GAN. The main contributions of this work are summarized as follows:

- We propose a brand-new asynchronous generative adversarial network called Async-GAN, which is simple, general (model-agnostic), and free of the drawbacks of existing solutions for asymmetric unpaired image-to-image translation.
- Using very basic base models, we show that Async-GAN can achieve currently state-of-the-art performance on various representative and challenging asymmetric image-to-image translation tasks. Our method can be extended to various models and achieve a large performance gain.

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

II. Related Work

A. Asymmetric Domain Translation

Recently, domain translation between two domains has drawn a lot of attention [26]–[30], which aims to bridge the knowledge between the source domain and the target domain. Ideally, the two domains have the same amount of information entropy, and consequently, dual learning is a general and elegant solution to handle domain translation. However, in most cases, the two domains are asymmetric, which indicates one domain is information-poor and another is relatively information-rich. The "poor-to-rich" domain translation is more difficult has wide applications: deblurring [31], [32], super-resolution [33], [34], colorization [2], [35], and so on. It is extremely challenging to obtain one precise "poor-to-rich" translation function for dual learning-based methods. Dou *et al.* [20] addressed this problem for the NIR-to-RGB face image translation and combined some additional specially-designed constraints to alleviate the information asymmetry problem. Li *et al.* [25] combined additional latent variables to synthesize samples with more diversity. However, it mainly performs experiments on low-resolution images and the synthesized images are still fuzzy. Chang *et al.* [24] adopted reference-guided generation scheme to generate images with target representation from the reference image. Besides, [22], [23], [36] elaborately designed asymmetric network architectures for various tasks. Tang *et al.* [22] designed two different network architectures for the translation (with more network capacity) and reconstruction functions. Fu *et al.* [23] conduct a geometry-consistent loss function for one-sided unsupervised domain mapping. These methods mainly focus on introducing extra data-dependent and task-specific conditions and fail to provide a general solution. It is also laborious to design different asymmetric architectures and search for the optimal hyper-parameters for different tasks. To tackle this problem, we propose a novel asynchronous generative adversarial network for unpaired asymmetric domain translation, which combines an innovative pseudo paired training strategy to boost the domain translation performance.

B. Unpaired Image-to-Image Translation

The unpaired image-to-image translation problem aims to translate the image with source representation to the required image with target representation without paired training data. The Cycle-GAN methods [4]–[6], [13], [37]–[41] become popular, which adopts the cycle-consistency loss to preserve the content information. Two separate functions (F and G) are responsible

for the forward translation and backward reconstruction, respectively. However, these methods are limited to performing image translation between two domains. To conduct the unpaired image translation among multiple domains, Choi *et al.* [15] extended the cycle-consistency and introduced one additional classification loss. With the identification of whether domain the image comes from, StarGAN [15] and STGAN [42] can synthesize images with multiple different representations using only one generator and discriminator. Besides, to boost the diversity of the synthesized samples of the generative model, the disentanglement methods such as MUNIT [5], DRIT [6] were developed by projecting the visual images to ideal orthogonal "style" and "content" domain. By swapping the style codes of images from two different domains, [5], [6] could synthesize a translated image with the target style by specifying a reference image. However, these methods commonly are restricted for low-resolution image translation and fail to synthesize plausible detailed outputs without full supervision. The "poor-to-rich" translation such as deblurring [31], [32] and super-resolution [33], [34], is also a big challenge due to the information asymmetry. For a real application, we require the model with a stronger translation ability to achieve high-resolution unpaired image-to-image translation. To enhance the unpaired image translation quality, we propose a novel pseudo paired training by introducing the intermediate samples. Our Async-GAN can achieve current state-of-the-art translation performance, and even are competitive with the translation methods with full supervision.

C. Strategies for Information Asymmetry

To handle the information asymmetry between domains, there have been several attempts explored. In detail, they aim to learn a high-capacity G ("poor-to-rich") and low-capacity F ("rich-to-poor") through multiple ways: 1) Making G bigger compared to F proposed in [22]. Tang *et al.* [22] proposed a stronger network with much more parameters to learn the "poor-to-rich" translation, which has inevitably induced high computational costs and inference time. 2) Training G more often compared to F or using a balanced learning rate for F and G. To avoid the mode collapse, the training balance between F and G is elaborately explored in [43], [44]. Different learning rates for F and G could be explored for asymmetric unpaired image-to-image translation. However, this thread of work does not utilize information asymmetry to obtain better results and they are still required to deliberately design the hyper-parameters to balance the training of F and G for various tasks. 3) Introducing extra constraints [23] or special guidance [18], [24], [25] for the "poor-to-rich" translation. Considering these attempts are data-dependent and task-specific, we propose a general framework to perform asynchronous learning. A novel modification of the unpaired translation model (utilize the frozen rich-to-poor translation to boost the training of poor-to-rich) does work quite well here and outperforms several recent methods. Furthermore, different from the above-mentioned methods, the proposed asynchronous training strategy could make good use of the information asymmetry and formulate self-generated full supervision to lead to better "poor-to-rich" translation performance.

III. The Proposed Approach

A. Overview Framework

Let $\mathbb{X}$ and $\mathbb{Y}$ denote two information-asymmetric image domains: an information-rich domain and an information-poor domain, respectively. Image-to-image translation from $\mathbb{Y}$ to $\mathbb{X}$ (*i.e.,* "poor-to-rich" translation) is a very challenging problem due to the large information gain demand (e.g. translating a highly abstract segmentation map to a natural and detailed RGB image), especially when the training data is unpaired, namely, there is no sample correspondence between $\mathbb{X}$ and $\mathbb{Y}$ when they stand for the training data. Suppose $x \in \mathbb{X}$ and $y \in \mathbb{Y}$ are two arbitrary samples, in the unpaired setting, x and y are not a pair. However, we can build pseudo sample pairs once the reliable domain-to-domain translation functions are learned. Suppose that $F(x) = y_t : \mathbb{X} \to \mathbb{Y}$ is a function transferring samples from $\mathbb{X}$ to $\mathbb{Y}$, *i.e.,* a "rich-to-poor" translation function, we can treat (x, y_t) or equivalently (y_t, x) as a pseudo sample pair if F is reliable, which also means that y_t can be regarded as a pseudo sample in domain $\mathbb{Y}$, or equivalently $\tilde{\mathbb{Y}} \approx \mathbb{Y}$ if $y_t \in \tilde{\mathbb{Y}}$. Here, $\tilde{\mathbb{Y}}$ is the intermediate domain which is close to domain $\mathbb{Y}$. Similarly, we can have $G(y) = x_t : \mathbb{Y} \to \mathbb{X}$ server as the "poor-to-rich" translation function which transforms samples from $\mathbb{Y}$ to $\mathbb{X}$. Unlike F, G is much harder to learn as mentioned earlier, so it is generally less reliable than F and thus we don't use it to generate pseudo sample pairs. To make $\tilde{\mathbb{Y}}$ as close to $\mathbb{Y}$ as possible, the widely used adversarial loss is adopted:

$$\begin{aligned}&\mathcal{L}_{adv}(F, D_y)\\&= \mathbb{E}_y[\log D_y(y)] + \mathbb{E}_{y_t}[\log(1 - D_y(y_t))], \text{with } y_t = F(x),\end{aligned} \tag{1}$$

where D_y is the discriminator of the information-poor domain $\mathbb{Y}$. The discriminator outputs the probability of the sample ($0 \leq D_y(y) \leq 1$, or $0 \leq D_y(y_t) \leq 1$) belonging to $\mathbb{Y}$. Similarly, we can also adopt the adversarial loss $\mathcal{L}_{adv}(G, D_x)$ for G and D_x, where D_x is the discriminator of the information-rich domain $\mathbb{X}$.

Since x and y are unpaired, the cycle-consistency loss $\mathcal{L}_{cyc}$ [4] is adopted to link F and G:

$$\mathcal{L}_{cyc}(F, G) = \mathbb{E}_x[\|G(F(x)) - x\|_1] + \mathbb{E}_y[\|F(G(y)) - y\|_1], \tag{2}$$

which makes F and G tied to each other so that they can help each other in the optimization (actually, it is more likely that F has to help G in the synchronous learning with $\mathcal{L}_{cyc}(F, G)$ due to the information-asymmetry). In addition to such synchronous learning loss which treats F and G equally during their synchronous optimization, our Async-GAN pays more attention to G by using frozen F to generate pseudo sample pairs for G after F becomes reliable enough, as shown in Algorithm.1 and Fig. 2. To make F reliable enough for pseudo sample pair generation, Async-GAN follows the synchronous learning strategy in the first half of its training epochs by using a synchronous total loss $\mathcal{L}_{Sync}$, for which a simple example is the typical Cycle-GAN loss:

$$\mathcal{L}_{Sync}(F, G, D_x, D_y)$$

$$= \mathcal{L}_{adv}(F, D_y) + \mathcal{L}_{adv}(G, D_x) + \lambda \mathcal{L}_{cyc}(F, G), \quad (3)$$

where λ is a super-parameter for loss balancing. We expect that after a large enough number of training epochs (*i.e.*, $\frac{N}{2}$ where N is the total number of training epochs) F can become reliable enough. Then, we employ the asynchronous learning strategy to do joint optimization (synchronously optimizing both F and G) and partial optimization (optimizing G only, with F frozen) alternately, in small cycles with a period of T epochs. In the first half of each cycle (*i.e.*, the first $\frac{T}{2}$ epochs) we use the synchronous loss $\mathcal{L}_{Sync}$ to train the model for joint optimization. In the second half (*i.e.*, the other $\frac{T}{2}$ epochs), we freeze F, which is denoted by $\bar{F}$, and use it for generating a pseudo sample pair $(\bar{y}_t, x)$ where $\bar{y}_t = \bar{F}(x)$ for each $x \in \mathbb{X}$, and then use such pseudo sample pairs for training G and D_x in a paired image-to-image translation manner, with the following partial loss:

$$\begin{aligned}\mathcal{L}_{Partial}(G, D_x) &= \mathcal{L}_{adv}(G, D_x) + \gamma \mathbb{E}_x[\|G(\bar{y}_t) - x\|_1] \\ &= \mathbb{E}_x[\log D_x(x)] + \mathbb{E}_{x_t}[\log(1 - D_x(x_t))] + \gamma \mathbb{E}_x[\|G(\bar{y}_t) - x\|_1]\end{aligned} \quad (4)$$

where γ is the weighting factor to balance the adversarial loss and the pixel-level content loss. With the adoption of the asynchronous learning, we can gradually promote the image quality of synthesized samples from the "intermediate domains". Thus, we can better formulate the full supervision for the more difficult translation G, which could result in better translation performance than symmetric solutions. We follow the default setting of the vanilla CycleGAN and set $\lambda = 10$ and $\gamma = 10$ in our all experiments to provide a fair comparison with CycleGAN.

B. Pseudo Paired Training

To better clarify the proposed method, our pseudo pair generation contains two main procedures: **pseudo pair generation** and **full-supervised learning**. At the former stage, we formulate the pseudo pairs: $(\tilde{y}_t = F(x), x)$. Note, the parameters of F are frozen at this stage and F is only responsible to synthesize corresponding translated images. The latter supervised training stage optimizes G through the full supervision between $G(\tilde{y}_t)$ and x. We formulate the self-generated pseudo images to provide the "poor-to-rich" translation model G with a strong constraint. To be noted, F and the corresponding domain-specific discriminator D_x are frozen during this training stage. With only a one-side translation function is optimized, the model tends to generate more accurate and precise detailed outputs. On the other hand, by providing direct pixel-level supervision, we can effectively reduce the uncertainty and noise when optimizing the model. The proposed training strategy of our Async-GAN is similar to the "pseudo label" proposed in [45], [46]. The pseudo sample pair $(\tilde{y}_t = F(x), x)$ could provide a full pixel-level supervision for the harder translation G. Through this, we can utilize the information asymmetry for better "poor-to-rich" translation performance.

Algorithm 1: Async-GAN.

Input: N (total No. of epochs), T (No. of epochs in each alternative optimization cycle), n_x (No. of training samples in $\mathbb{X}$) and n_y (No. of training samples in $\mathbb{Y}$).
Output: Model parameters $\theta = \{\theta_F, \theta_G, \theta_{D_x}, \theta_{D_y}\}$.

```
for n = 1, 2, ..., N/2 do
  for i, j = 1, 2, ..., min(n_x, n_y) do
    Random sample x_i, y_j from X and Y,
    Optimize all the parameters θ with L_Sync (Eq. 3),
  end for
end for
for n = N/2 + 1, N/2 + 2, ..., N do
  k = n%T,
  if k ≤ T/2 then
    for i, j = 1, 2, ..., min(n_x, n_y) do
      Random sample x_i, y_j from X and Y,
      Optimize all the parameters θ with L_Sync
        (Eq. 3),
    end for
  else
    for i = 1, 2, ..., n_x do
      Random sample x_i from X,
      Optimize θ_G and θ_{D_x} using pseudo sample
        pair (ȳ_t^i, x_i) and partial loss L_Partial (Eq. 4),
        with frozen θ_F and θ_{D_y}.
    end for
  end if
end for
```

IV. Experiments

A. Datasets

For datasets with image pairs, the information-rich domain and information-poor domain have the same number of images. We provide the number of detailed training samples from the two domains.

Cityscapes [47] is a large-scale dataset consisting of diverse urban street scene videos captured across 50 different cities at varying times of the year. It provides the ground truths for several vision tasks including semantic segmentation. We adopt its RGB image and semantic label map pairs to perform the translation from the semantic label maps to the photo images and utilize the train/val split of the raw dataset for our experiments.

Foggy Cityscapes [21] is a simulated synthetic foggy dataset with three different levels of visibility: 150 m, 300 m and 600 m visibility. For this dataset, we choose clear images from Cityscapes dataset [47] and foggy images with 150 m visibility following the original split. 2,975 clear and the corresponding foggy image pairs are applied for training and 500 clear and foggy image pairs for evaluation.

BDD100 K [48] is a large scale high-resolution autonomous driving dataset containing 100,000 video clips from various cities and under different conditions. It provides various annotations for a keyframe from each video clip, including object bounding boxes and pixel-wise semantic segmentation masks,

resulting in a total of 100,000 annotated images. We only adopt pixel-wise semantic annotation for evaluation. We reorganized this dataset for night-to-day tasks according to the annotation of the daytime data and obtained a 27,971/3,929 train/test split for night images and a 36,728/5,258 train/test split for day images. To be noted, all the images from BDD100 K are unpaired. For the evaluation stage, we also measure the translation performance using the video sequences.

APDrawing [49] dataset includes 490 high-resolution photo-portrait image pairs, all the portrait images are from professional artistic drawings. We apply the evaluation protocol of [49] with a training/test split ratio of 6:1.

Facades dataset and **Aerial Map** dataset also contain the natural images with detailed information and the corresponding semantic label images. In Facades [4] dataset, each pair is made up of an image of a semantic label map and a photo of the same building. We use 400/106 pairs of train/test samples following CycleGAN [4]. As for the Aerial Map dataset, the train/test split is 1096/1098.

B. Implementation Details and Comparisons

Considering the homogeneous training flowchart between Async-GAN and Cycle-GAN based methods, we choose CycleGAN [4] for our baseline. To achieve high-resolution unpaired image-to-image translation, we extend the depth of the generator architecture of the CycleGAN method to perform image synthesis. Besides, the depth of the discriminators for checking whether the images are realistic also increases. In detail, we add one downsampling module and one upsampling module for the generators. As for the discriminator, one additional downsampling module is adopted compared with the original CycleGAN. For performing unpaired training with symmetric structure, two generators of the same architecture are designed to do the bidirectional translations. Cycle-consistency loss [4] is also adopted to guarantee content information. The multi-scale discriminator architecture is integrated to enhance the quality of generated images. In general detail, the auto-encoder generator structure is defined as:

Encoder:

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

Decoder:

$-Res256 - Res256 - Res256 - DI256F4 - DI128F4 - D64F7 - D3$

$CImFn$ means the Convolution-InstanceNorm-ReLU layer with m $n \times n$ convolution filters while $DImFn$ means the Deconvolution-InstanceNorm-ReLU layer with m $n \times n$ convolution 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.

We compare the proposed Async-GAN with the state-of-the-art methods designed for both paired and unpaired image-to-image translation. The chosen unpaired methods include CycleGAN [4], MUNIT [5], DRIT [6], StarGAN-V2 [9], and the paired methods are Pix2pix [2], BicycleGAN [50] and Pix2pixHD [3] if paired samples exist for the adopted datasets. We follow the official instructions of those methods and make a fair setting for comparison. Besides, to eliminate the influence of the image resolution for the training of current unpaired methods, we have designed experiments under different image resolutions to provide an intuitive comparison. To save computation costs and provide a direct comparison, we perform two representative unpaired training methods: CycleGAN [4] and MUNIT [5] under different image resolution settings. The original CycleGAN and MUNIT both perform unpaired image translation under the image resolution 256×256. We modified the input resolution of CycleGAN and MUNIT and designed: CycleGAN-512 (with image resolution 512×512); CycleGAN-1024 (with image resolution 1024×512); MUNIT-512 (with image resolution 512×512) and MUNIT-1024 (with image resolution 1024×512). Considering MUNIT and DRIT share similar architecture and design, we do not modify DRIT under our setting. Similarly, we add one downsampling and upsampling module for the generators of CycleGAN-512 and MUNIT-512, and one downsampling module for the discriminators of CycleGAN-512 and MUNIT-512. To be noted, CycleGAN-512 and CycleGAN-1024 share the same network architecture while the computational cost is different. For qualitative comparison in the following experiment parts, we only illustrate the synthesized results under 1024×512 image resolution for these two methods.

C. Evaluation Metrics

FID is proposed in [51] to measure the distance between the real sample distribution and the synthesized sample distribution. A lower FID score indicates stronger image translation performance. We adopt the FID score to measure the image translation performance in our paper.

LPIPS (Perceptual Image Patch Similarity [52]) computes the perceptual similarity between two images based on image patches. A lower LPIPS score indicates more perceptual similarity between two images. We compute this metric between the translated outputs and corresponding ground-truth images to measure the image translation ability.

Image Quality Metrics. Traditional image quality metrics include **PSNR** and **SSIM** are also adopted. PSNR (Peak Signal-to-Noise Ratio) [53] can roughly evaluate image quality independently, and the higher PSNR indicates the better performance. SSIM (Structural Similarity Index Measure) [53] measures the structural similarity between output and target images, and a higher SSIM denotes better structural similarity.

Semantic segmentation performance is also chosen to measure the image synthesis quality, we adopted a Deeplab-v3 model[1] pre-trained on Cityscapes dataset [47], performing semantic segmentation on the translated outputs based on the evaluation scripts.[2] We compute the per-class pixel accuracy (denoted as "Per-class acc.") and the Intersection-Over-Union (IoU) between the outputs and the ground-truths. We report the mean IoU (mIoU) and the pixel accuracy of all categories. Please note

[1][Online]. Available: https://github.com/srihari-humbarwadi/DeepLabV3_Plus-Tensorflow2.0

[2][Online]. Available: https://github.com/mcordts/cityscapesScripts

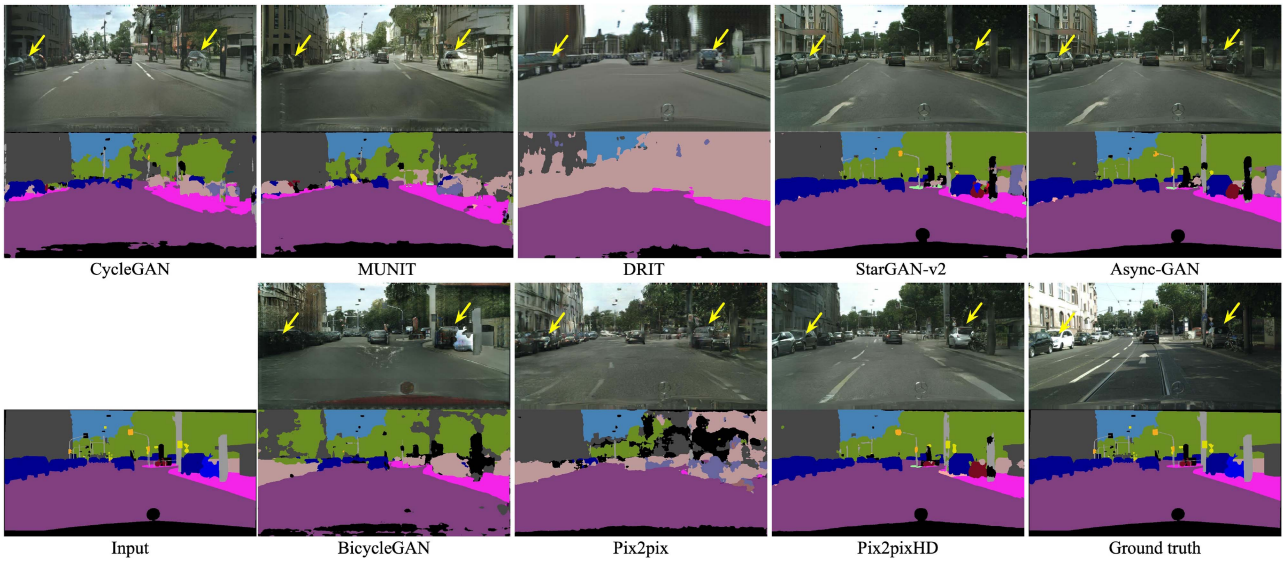

Fig. 3. The visual comparison of different methods for label-to-photo translation on Cityscapes dataset. We exhibit the translated results and their corresponding segmentation outputs at the same time. Details where the yellow arrows pointed to are worth of attention. For CycleGAN and MUNIT, we illustrate the qualitative results of CycleGAN-1024 and MUNIT-1024, respectively.

TABLE I
QUANTITATIVE COMPARISON OF LABEL-TO-PHOTO TRANSLATION ON CITYSCAPES DATASET. THE SYMBOL ↑ (↓) INDICATES THAT THE LARGER (SMALLER) THE VALUE, THE BETTER THE PERFORMANCE. METHODS ABOVE THE BOLD LINE PERFORM THE UNPAIRED TRAINING, WITH THEIR BEST RESULTS SHOWN IN **BOLD**, AND METHODS BELOW THE BOLD LINE ARE TRAINED USING PAIRED DATA, WITH THEIR BEST RESULTS SHOWN IN *ITALIC*. TO BE NOTED, WE EVALUATE TRADITIONAL EVALUATION METRICS AND THE SEGMENTATION PERFORMANCE OF ALL THE METHODS AT 1024×512 AND 2048×1024 RESOLUTIONS, SEPARATELY (WE ENLARGE THE SYNTHESIZED IMAGES TO THE SAME SIZE BY APPLYING BILINEAR INTERPOLATION)

Method	Unpaired	Resolution	SSIM↑	PSNR↑	LPIPS↓	FID↓	Per-class acc.↑	mIoU↑
CycleGAN-256	√	256 × 256	0.8347	14.87	0.5451	77.11	25.1	8.3
CycleGAN-512	√	512 × 512	0.8328	15.03	0.5134	84.93	36.3	22.1
CycleGAN-1024	√	1024 × 512	**0.8415**	15.12	0.4617	71.64	42.6	27.7
MUNIT	√	256 × 256	0.7744	13.80	0.5481	71.85	28.8	17.1
MUNIT-512	√	512 × 512	0.8156	14.61	0.4804	78.25	35.6	21.6
MUNIT-1024	√	1024 × 512	0.8254	14.98	0.4615	84.16	40.6	26.5
DRIT	√	256 × 256	0.7714	13.95	0.5584	115.3	25.7	11.6
StarGAN-v2	√	512 × 512	0.8415	15.56	0.4417	78.73	55.6	41.7
Async-GAN	√	1024 × 512	0.8396	**15.67**	**0.4374**	**72.10**	**58.1**	**47.7**
Pix2pix	×	256 × 256	*0.8588*	16.02	0.4991	85.60	29.1	17.3
BicycleGAN	×	256 × 256	0.7960	14.41	0.5924	153.5	31.2	18.9
Pix2pixHD	×	1024 × 512	0.8528	*16.09*	*0.4366*	*59.38*	*67.4*	*58.6*

that the same pre-trained Deeplab-V3 model on Cityscapes is chosen to conduct a fair comparison.

D. Comparison With the State-of-The-Art

Label-to-Photo Translation: Semantic labels provide the pixel-level annotation of visual images, which highly abstract the visual category information. Compared with the semantic labels, the raw visual images contain much richer details, *e.g.*, various colors, and complex texture. In general detail, the different shapes and types of cars share the same instance. The asymmetric translation from the semantic labels to the visual photo images is extremely challenging without the paired training. We perform label-to-photo translation tasks on Cityscapes dataset [47] under both the paired and unpaired training settings. The visual synthesized results of all the methods are shown in Fig. 3. To better illustrate the visual translation performance, we exhibit the corresponding semantic segmentation outputs of all the translated outputs using a pre-trained Deeplab-V3 model on Cityscapes dataset [47]. The better segmentation performance can indirectly indicate precise image translation performance.

From both the visual synthetic results and segmentation outputs, our Async-GAN outperforms all the unpaired methods by a large margin. Our Async-GAN can synthesize cars and pedestrians precisely and generate clearer boundary sketches of the adhesive cars. Even compared with the fully supervised paired methods: Pix2pix [2] and BicycleGAN [50], our method can also obtain better synthesis performance. Our method and the high-resolution paired training method Pix2pixHD [3] that can be regarded as the upper bound of our Async-GAN are head to head. As for quantitative comparison shown in Table I, our method achieves remarkable segmentation performance (both the pixel accuracy and mIoU), which indicates that Async-GAN can achieve the best asymmetric unpaired translation even without paired training samples. As for the comparison with AsymGAN [25], we did not conduct experiments due to the

Fig. 4. The qualitative comparison of different methods Foggy Cityscapes [21] dataset. For CycleGAN and MUNIT, we illustrate the qualitative results of CycleGAN-1024 and MUNIT-1024, respectively. Best viewed in color.

TABLE II
QUANTITATIVE COMPARISON OF DEFOGGING TASK ON FOGGY CITYSCAPES [21] DATASET. METHODS ABOVE THE BOLD LINE PERFORM THE UNPAIRED TRAINING, WITH THEIR BEST RESULTS SHOWN IN **BOLD**, AND METHODS BELOW THE BOLD LINE ARE TRAINED USING PAIRED DATA, WITH THEIR BEST RESULTS SHOWN IN *ITALIC*. TO BE NOTED, WE EVALUATE TRADITIONAL EVALUATION METRICS AND THE SEGMENTATION PERFORMANCE OF ALL THE METHODS AT 1024×512 AND 2048×1024 RESOLUTIONS, SEPARATELY

Method	Unpaired	Resolution	SSIM↑	PSNR↑	LPIPS↓	FID↓	Per-class acc.↑	mIoU↑
CycleGAN	√	256 × 256	0.8413	20.16	0.4136	78.98	56.8	32.4
CycleGAN-512	√	512 × 512	0.8604	22.76	0.3712	67.54	63.1	38.9
CycleGAN-1024	√	1024 × 512	0.8656	23.24	0.3604	71.87	69.3	44.1
MUNIT	√	256 × 256	0.8517	21.25	0.3917	84.62	47.8	28.7
MUNIT-512	√	512 × 512	0.8692	22.42	0.3667	73.61	69.2	42.1
MUNIT-1024	√	1024 × 512	0.8701	23.41	0.3535	64.62	71.1	46.2
DRIT	√	256 × 256	0.8542	22.19	0.3831	92.98	51.1	31.1
StarGAN-v2	√	512 × 512	0.8812	26.12	0.3426	**59.75**	79.1	54.9
Async-GAN	√	1024 × 512	**0.8891**	**26.48**	**0.3313**	62.10	**85.8**	**59.3**
Pix2pix	×	256 × 256	0.8516	23.41	0.3623	85.60	67.8	42.5
BicycleGAN	×	256 × 256	0.8443	21.71	0.3965	78.87	61.2	38.5
Pix2pixHD	×	1024 × 512	*0.9051*	*27.18*	*0.3414*	*59.38*	*84.9*	*57.9*

official implementation and configuration of AsymGAN are not released, so it is hard to conduct a fair comparison under all our settings. Instead, we can provide a quantitative comparison with the reported results on the Cityscapes dataset. As reported, AsymGAN achieved 24.5 mIOU and 31.4 Per-class accuracies, while our Async-GAN obtained 47.7 mIOU and 58.1 Per-class accuracies, which are nearly doubled compared with AsymGAN. Our Async-GAN can generate high-resolution images with high image quality while AsymGAN mainly focuses on improving the diversity of generating low-resolution images by combining additional latent variables.

Defogging: We also conduct comprehensive experiments on the fogginess removal task. Due to the existence of dense fog, some content information of the foggy image is missed, which results in the information asymmetry between the clear and foggy images. Following the same experimental setup of the above label-to-photo translation, we perform defogging on the Foggy Cityscapes dataset [21], with qualitative results shown in Fig. 4. The quantitative comparison is also reported in Table II. As illustrated, MUNIT and CycleGAN can only generate fuzzy outputs while some detailed information has been lost after the image translation. Our Async-GAN can effectively remove the dense fog of the input foggy images and synthesize images with reasonable content representations. Besides the visual quality measurement, we also evaluate the semantic segmentation performance to directly measure the defogging performance. The pixel accuracy and mIoU scores of different methods are also provided in Table II. From the comparison, our method can still achieve the best semantic segmentation performance. To be noted, the mIoU score of the original foggy images is only 25.6. By comparing the mIoU scores of the original foggy images and translated images, our Async-GAN could effectively enhance the image quality and lead to the performance gain for the downstream semantic segmentation task.

Night-to-Day Translation: We perform the night-to-day translation experiments on the BDD100 K dataset. Usually, there is more complex information in daytime images, including more pedestrians, meaningful text information on the billboards, and dynamic objects. Vice versa, there are less informative parts in images captured at night and it is also difficult for both the vision system and humans to achieve effective recognition. The reflection of glasses and road surface and low illumination constitute two main factors that lead to the poor performance of deep convolutional neural networks. The daytime images have relatively richer information than nighttime images. We apply the proposed asynchronous generative adversarial network to the night-to-day translation. Since there are no paired samples in BDD100 K dataset [48], we do not implement the paired image-to-image translation methods. The visual comparative translation outputs of all unpaired methods are shown in Fig. 5. Besides, we also show the segmented outputs of the translated daytime images and the raw nighttime images. Better

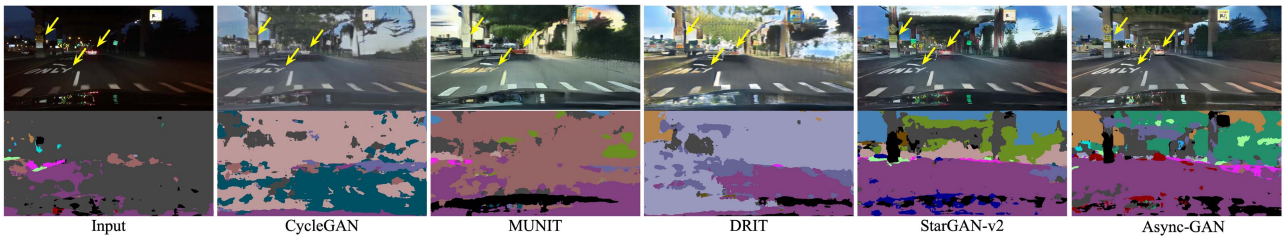

Fig. 5. The visual comparison of different methods on BDD100 K dataset, with both the images and their corresponding segmentation results shown for one representative samples. For CycleGAN and MUNIT, we illustrate the qualitative results of CycleGAN-1024 and MUNIT-1024, respectively.

TABLE III
QUANTITATIVE COMPARISON OF NIGHT-TO-DAY TRANSLATION ON BDD100 K DATASET [48]. WE EVALUATE THE SEGMENTATION PERFORMANCE OF ALL METHODS AT 1280 × 720 RESOLUTION

Method	Resolution	FID↓	Per-class acc.↑	mIoU↑
Original	1280 × 720	101.2	39.8	14.1
CycleGAN-256	256 × 256	80.82	19.4	10.3
CycleGAN-512	512 × 512	76.45	37.8	14.6
CycleGAN-1024	1024 × 512	73.87	41.4	16.7
MUNIT-256	256 × 256	61.83	19.7	9.60
MUNIT-512	512 × 512	59.76	34.6	13.2
MUNIT-1024	1024 × 512	61.83	38.5	14.0
DRIT	256 × 256	68.99	34.5	15.8
StarGAN-v2	512 × 512	51.67	45.4	25.6
Async-GAN	1024 × 512	**41.92**	**65.1**	**31.6**

segmentation outputs indirectly indicate a stronger translation ability. As exhibited, our method can achieve "poor-to-rich translation" precisely without paired training. The translation function G can be better optimized by our pseudo paired training. The text information on the billboard (pointed by the yellow arrows) has been preserved after the translation, while other unpaired translation methods fail to synthesize meaningful text information after translation. At the same time, our method can also enhance the weak information (the boundary of the adhesive cars) of the nighttime images and boost the segmentation performance (from 14.1 to 31.6 on mIoU score). As shown by the quantitative comparison of different methods reported in Table III, our method achieves the best translation performance and outperforms other unpaired image translation methods by a large margin. Please check more night-to-day translation results on our website.

Portrait-to-Photo Translation: The portrait images provide a highly abstracted characteristic representation of one person while the photo images cover more detailed color and fine-grained characters information. We implement the portrait-to-photo translation on the collected paired photo-portrait dataset [49]. Following the same setting, we compare both unpaired and paired translation methods. Similarly, we also perform CycleGAN-512 (512 × 512) and MUNIT-512 (512 × 512). The visual comparison of all the different methods is shown in Fig. 6. As shown, our Async-GAN can generate portrait images with reasonable facial features and less dirty color blocks. We also report the quantitative comparison except for the segmentation results in Table IV, Async-GAN achieved the best translation performance among all the unpaired methods. Even when compared with the paired methods, our method can surpass

Fig. 6. The visual comparison of different methods on APDrawing dataset [49].

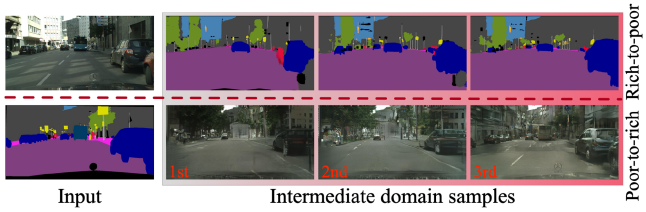

Fig. 7. The intermediate domain samples of both rich-to-poor and poor-to-rich directions.

Pix2pix and BicycleGAN by a large margin. Our Async-GAN and Pix2pixHD [3] methods are neck and neck.

E. Ablation Study

Intermediate Domains: For a better understanding of our intermediate domains, we exhibit translated results of intermediate domains in Fig. 7. The plausible rich-to-poor results can provide better pseudo paired supervision for poor-to-rich translation. With asynchronous training, we can improve the poor-to-rich translation ability iteratively. Conversely, we can also boost the rich-to-poor translation based on a stronger poor-to-rich mapping function. And the quantitative comparison of the different intermediate domains is shown in Table V. As shown in this table, we can enhance the image synthesis quality (from 93.61 to 72.1 in terms of FID score) through our asynchronous training strategy.

Selection of T: To explore the sensitiveness of our method to T, we have conducted the experiments and the experimental results are shown in Table VI. As reported, the combination of $(N = 120, T = 20)$ could achieve the best translation performance on the Cityscapes dataset. Furthermore, we have to

TABLE IV
QUANTITATIVE COMPARISON OF PORTRAIT-TO-PHOTO TRANSLATION ON APDRAWING [49] DATASET

Method	Unpaired	Resolution	SSIM↑	PSNR↑	LPIPS↓	FID↓
CycleGAN-256	√	256×256	0.8329	12.62	0.4578	125.4
CycleGAN-512	√	512×512	0.8387	13.21	0.4315	111.5
MUNIT-256	√	256×256	0.8199	12.85	0.4636	139.6
MUNIT-512	√	512×512	0.8376	13.16	0.4227	105.2
DRIT	√	256×256	0.8125	12.08	0.4875	144.8
StarGAN-v2	√	512×512	0.8494	**13.64**	0.4175	**98.56**
Async-GAN	√	512×512	**0.8524**	13.54	**0.4165**	121.1
Pix2pix	×	256×256	0.8504	*13.48*	0.4879	151.4
BicycleGAN	×	256×256	0.8307	11.97	0.4952	182.4
Pix2pixHD	×	512×512	*0.8553*	13.47	*0.4346*	*135.4*

TABLE V
QUANTITATIVE COMPARISON OF THE SYNTHESIZED SAMPLES OF THE DIFFERENT INTERMEDIATE DOMAINS

Method	FID↓	SSIM↑	PSNR↑	LPIPS↓
1st	93.61	0.7923	14.92	0.5367
2nd	82.50	0.8178	15.35	0.4812
3rd	**72.10**	**0.8396**	**15.67**	**0.4374**

TABLE VI
QUANTITATIVE FID AND MIOU COMPARISON OF LABEL → PHOTO TRANSLATION TASK ON CITYSCAPES DATASET UNDER VARIOUS SETTINGS

Method	FID↓	mIoU↑
$(N = 100, T = 5)$	76.3	41.9
$(N = 100, T = 15)$	78.1	45.1
$(N = 100, T = 30)$	82.8	39.2
$(N = 100, T = 10)$	**72.1**	**47.7**

admit that the selection of T requires some empirical prior and depends on the datasets. Various datasets should have different optimal hyper-parameter combinations.

General Strategy: Our method is a general and elegant training strategy, which means that it can be extended to other methods and achieve performance gain. In addition to the baseline architecture that our Async-GAN adopts, we also incorporate our asynchronous training strategy with some other representative unpaired image-to-image translation methods: CycleGAN and MUNIT, denoted by Async-CycleGAN and Async-MUNIT, respectively. The quantitative result comparisons are reported in Table VII. The asynchronous strategy significantly promotes all the based methods on the segmentation performance of the translated outputs although it may hurt a little bit on one or two other image quality metrics, which shall not be considered a big deal as these metrics alone are not so indicative. The consistent improvements demonstrate that the proposed asynchronous generative adversarial learning strategy is general and model-agnostic for the asymmetric unpaired image translation. Besides, our Async-GAN can also achieve the best translation result among all the settings as a result of adopting the high-resolution translation backbone.

Compatibility with multimodal translation: There is a tradeoff between image translation quality and image synthesis diversity. To explore the compatibility with the multimodal translation task, we propose to introduce the additional latent variables into our framework to perform multimodal image translation. In detail, referring to styleGAN [54], we sample a latent code z (a 128-dimensional vector) from the latent space then we inject the sampled z at the bottleneck. Then we aim to reconstruct the latent code $\hat{z}$ from the translated images and perform adversarial training among the latent space adopted in AsymGAN [25]. We have extended our Async-GAN to handle the multimodal label → photo translation task on the Cityscapes dataset. The qualitative results are shown in Fig. 8. By choosing different latent codes z, we can obtain various translation outputs with different appearance representations. To measure the image synthesis diversity, we adopt average LPIPS scores among all the samples proposed in [55]. The FID scores are also reported to measure the image synthesis performance. For better illustration, we also report the results of the specially designed BicycleGAN for the multimodal image translation task. The quantitative comparison is illustrated in Table VIII. As observed, with the help and guidance of the latent codes, our Async-GAN could generate much diverse image translation outputs and distributions that are generally closer to the real sample distribution indicated by a lower FID score. Besides, following the same experimental setting, we also perform experiments on label → photo on Facades dataset and report the quantitative comparison in Table VIII. From these results, we can see that our Async-GAN has good compatibility with multimodal translation by adopting some extra latent variables.

Rich-to-poor translation: Though our method is mainly designed for the poor-to-rich image translation, we have also conducted the corresponding "rich-to-poor" translation experiments to show that the proposed method could also promote the "rich-to-poor performance". The results are reported in Table IX and the quantitative scores of AsymGAN are also exhibited for reference. We report the per-pixel accuracy, per-class accuracy, and mean IoU scores of different classes. For better comparison, we also provide the quantitative results of two baseline methods (CycleGAN-1024 and MUNIT-1024). Our Async-GAN could achieve the best performance for both the photo → label and label → photo translation tasks. AsymGAN only achieved 74.9%, 27.6% and 21.6% while the proposed method got 81.6%, 44.7% and *32.6%* under the photo → label setting.

$\mathcal{L}_{Sync}$ and $\mathcal{L}_{Partial}$ dissection. In our Async-GAN, we optimize all the parameters θ with $\mathcal{L}_{Sync}$ in line 12 of Algorithm.1. The readers may concern why not only optimizing F and D_y

TABLE VII
QUANTITATIVE COMPARISON OF LABEL-TO-PHOTO TRANSLATION ON CITYSCAPES DATASET FOR DIFFERENT METHODS. TO BE NOTED, ALL THE METHODS ARE OPTIMIZED BASED ON UNPAIRED DATA. CYCLEGAN* AND MUNIT* DENOTE THE CORRESPONDING COUNTERPARTS WITH THE PROPOSED ASYNCHRONOUS LEARNING

Method	Asynchronous	Resolution	SSIM↑	PSNR↑	LPIPS↓	FID↓	Per-class acc.↑	mIoU↑
CycleGAN	×	256 × 256	0.8347	14.87	0.5451	77.11	25.1	8.3
CycleGAN*	√	256 × 256	0.8109	14.28	0.5102	65.51	31.6	18.3
MUNIT	×	256 × 256	0.7744	13.80	0.5481	71.85	28.8	17.1
MUNIT*	√	256 × 256	0.7729	14.17	0.4963	81.89	37.6	24.5
CycleGAN-1024	×	1024 × 512	0.8415	15.12	0.4617	71.64	42.6	27.7
Async-GAN	√	1024 × 512	0.8396	15.67	0.4374	72.10	58.1	47.7

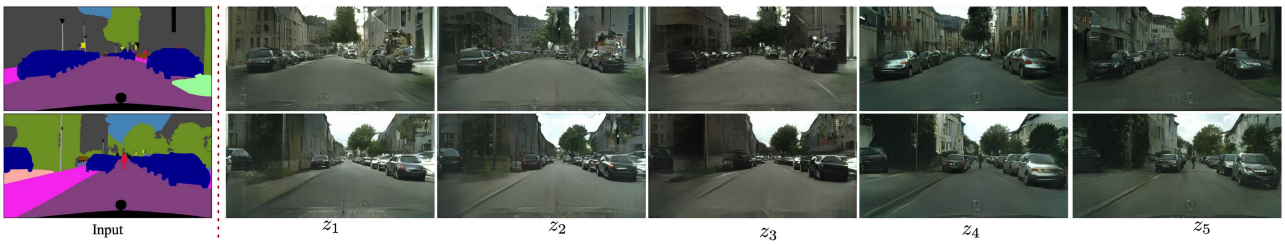

Fig. 8. The qualitative label → photo translation results by combining the additional latent codes z on Cityscapes dataset. z_1, z_2, z_3, z_4, z_5 indicate various sampled latent codes.

TABLE VIII
QUANTITATIVE FID (FOR IMAGE TRANSLATION PERFORMANCE MEASUREMENT, LOWER IS BETTER) AND LPIPS (FOR IMAGE SYNTHESIS DIVERSITY MEASUREMENT, HIGHER IS BETTER) COMPARISON OF LABEL → PHOTO TRANSLATION TASK ON CITYSCAPES AND FACADES DATASETS

Method	Dataset	FID↓	LPIPS↑
BicycleGAN	Cityscapes	89.42	0.1456
Async-GAN(w/o z)	Cityscapes	72.10	0.1087
Async-GAN(w/ z)	Cityscapes	**60.41**	**0.1521**
BicycleGAN	Facades	62.95	0.1543
Async-GAN(w/o z)	Facades	65.61	0.1402
Async-GAN(w/ z)	Facades	**51.24**	**0.1902**

since we only require a strong F, provide high-quality pseudo sample pair for G and D_x to compute $\mathcal{L}_{Partial}$ in line 19 of Algorithm 1. If we just optimize F and D_y based on $\mathcal{L}_{Sync}$, G and D_x could be regarded frozen (G is fixed). As above discussed, it is very challenging to obtain a reliable poor-to-rich translator G through the cycle-consistency based training. Thus, G may output the noisy output $G(y)$ and the reconstruction loss between $F(G(y))$ and y is also computed to optimize F. Since G is not as reliable as F, the training of F may be plagued by G and there would be performance degradation for the rich-to-poor translation. We conduct the experiments under the suggested setting and the qualitative translation results of $F(x)$ are reported in Fig. 9. We observe there is an obvious performance drop if we just optimize θ_F and θ_{D_y} with $\mathcal{L}_{Sync}$. Furthermore, The pixel-wise reconstruction part in $\mathcal{L}_{Partial}$ is also the key to providing the full supervision for the poor-to-rich translator based on the pseudo paired training samples. We have removed the reconstruction loss $\mathcal{L}^{rec}_{Partial} = \|G(F(x)) - x\|_1$ and performed corresponding experiments on the label → photo translation task on Cityscapes. There is huge performance degradation: the mIoU score dropped from 47.7 to 28.3). Thus, we can conclude that the reconstruction loss is a necessary part of the proposed Async-GAN.

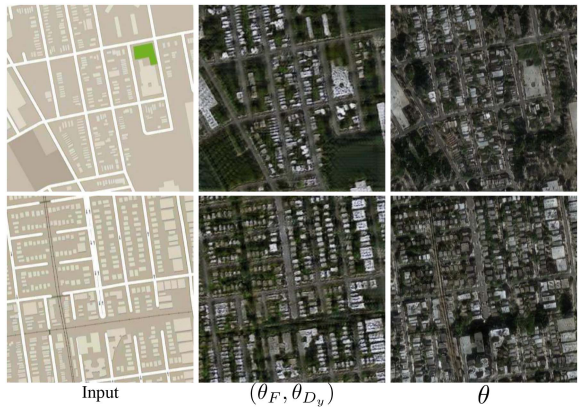

Fig. 9. The visual "poor-to-rich" translation output comparison of optimizing (θ_F, θ_{D_y}) and θ in line 12 of Algorithm 1, respectively.

Information Asymmetry: Besides verifying the effectiveness of our proposal on the translation between two asymmetric domains, we also investigate how it performs when the two domains are symmetric, such as summer ↔ winter [4], woman ↔ man [15]. We test CycleGAN and Async-CycleGAN on both tasks. The quantitative results under the FID metric are shown in Table X. Compared with CycleGAN, Async-CycleGAN only shows a marginal superiority which is hard to be considered as significant. Therefore, the power of Async-GAN comes from its exploration of the hidden energy inside the information asymmetry. The more asymmetry the amount of information in two visual image domains, the larger the improvement our Async-GAN can achieve compared with the vanilla Cycle-GAN methods.

F. More Results

To further demonstrate the effectiveness of our method, we also performed various translation tasks with information asymmetry and the proposed method could achieve performance gain

TABLE IX
QUANTITATIVE COMPARISON OF BOTH LABEL → PHOTO AND PHOTO → LABEL TRANSLATION TASKS ON CITYSCAPES DATASET

Methods	**Label→Photo**			**Photo→Label**		
	Per-pixel acc. ↑	Per-class acc. ↑	mIoU ↑	Per-pixel acc. ↑	Per-class acc. ↑	mIoU ↑
AsymGAN	79.4	31.4	24.5	74.9	27.6	21.6
CycleGAN-1024	81.1	42.6	27.7	79.4	38.1	30.1
MUNIT-1024	79.7	40.6	26.5	76.6	34.6	28.5
Async-GAN	**86.2**	**58.1**	**47.7**	**81.6**	**44.7**	**32.6**

TABLE X
QUANTITATIVE FID COMPARISON OF SUMMER ↔ WINTER AND WOMAN ↔ MAN TRANSLATION TASK. THE RED AND BLUE VALUES REPRESENT THE FORWARD AND BACKWARD TRANSLATION PERFORMANCE, SEPARATELY. LOWER RESULTS ARE BETTER

Task	summer↔winter	woman↔man
CycleGAN	56.26 / **62.47**	34.35 / 29.82
Async-CycleGAN	**51.64** / 64.86	**31.73** / **26.14**

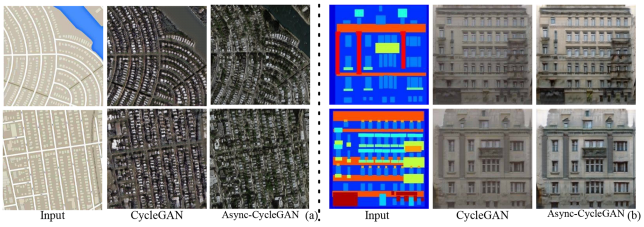

Fig. 10. The visual "poor-to-rich" translation outputs. (a) shows the results from the translation from aerial map to photo images while (b) exhibits the translation from label to facades. We also provide the translation results of vanilla CycleGAN to illustrate the difference. Best viewed in color and zoom in to check the details.

without introducing additional parameters or specifically designed loss functions.

Aerial map to photo: we first perform our Async-GAN on the Aerial map [2] dataset to achieve high-resolution translation (with 512×512 image resolution) from the aerial map to photo images shown in Fig. 10(a) following the official train/test split. The boundary of different objects is synthesized precisely. An intuitive comparison with the vanilla CycleGAN baseline is also provided. As illustrated, our method can generate more detailed information with fewer visual artifacts, which indicates the proposed method has a stronger translation ability to handle the "poor-to-rich" image translation task.

Facades dataset [56]: besides the translation from the aerial map to photo images, we also report the high-resolution translation results from label to facades are shown in Fig. 10(b). Similarly, compared with CycleGAN, our method can still generate more precise image outputs from the semantic label to the real frontal facades images.

Downstream digit classification: To demonstrate that the proposed method could serve for the downstream vision tasks, following the experimental setting of GcGAN [23],we conduct the MNIST ↔ SVHN and report the cross-domain digit classification results. CycleGAN could achieve 26.1% classification accuracy under the SVHN → MNIST setting while GcGAN has achieved 33.3%. In contrast, our method could achieve 38.8% classification accuracy, which has demonstrated the superiority of our method.

V. CONCLUSION

In this paper, we proposed a novel generic asynchronous generative adversarial network (termed Async-GAN) for asymmetric unpaired image-to-image translation (mainly focusing on the direction from the information-poor domain to the information-rich domain). Async-GAN utilizes the use of the information asymmetry between the domains to build gradually improving *intermediate domains* for generating pseudo paired samples from the relatively easier rich-to-poor translation and then uses these pseudo samples to aid the original unpaired data-based training through the full supervision. It is a model-agnostic (independent from backbone networks) framework that can be used for enhancing existing models for much better performance on unpaired image-to-image translation, without bringing any new model design efforts and extra parameters. Comprehensive experimental results on various tasks demonstrated the superiority of the proposed approach. We believe that our method can be extended for asymmetric domain translation between heterogeneous data (e.g. text and image, text and audio), which is left as our future work.

REFERENCES

[1] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.
[3] T.-C. Wang *et al.*, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8798–8807.
[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.
[5] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 172–189.
[6] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–51.
[7] Z. Zheng, Y. Wu, X. Han, and J. Shi, "ForkGAN: Seeing into the rainy night," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 155–170.
[8] Z. Zheng *et al.*, "Unpaired photo-to-caricature translation on faces in the wild," *Neurocomputing*, vol. 355, pp. 71–81, 2019.
[9] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8188–8197.
[10] Z. Zheng, Z. Yu, H. Zheng, Y. Yang, and H. T. Shen, "One-shot image-to-image translation via part-global learning with a multi-adversarial framework," *IEEE Trans. Multimedia*, vol. 24, pp. 480–491, 2021.
[11] J. Huang, J. Liao, and S. Kwong, "Semantic example guided image-to-image translation," *IEEE Trans. Multimedia*, vol. 23, pp. 1654–1665, 2021.
[12] H. Emami, M. M. Aliabadi, M. Dong, and R. B. Chinnam, "SPA-GAN: Spatial attention GAN for image-to-image translation," *IEEE Trans. Multimedia*, vol. 23, pp. 391–401, 2021.

[13] C. Wang *et al.*, "Discriminative region proposal adversarial networks for high-quality image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 770–785.
[14] J. Kim, M. Kim, H. Kang, and K. H. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *Proc. Int. Conf. Learn. Representation*, 2020.
[15] Y. Choi *et al.*, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8789–8797.
[16] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.
[17] C. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.
[18] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "Unpaired portrait drawing generation via asymmetric cycle mapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8217–8225.
[19] R. Liu, Q. Yu, and S. Yu, "Unsupervised sketch-to-photo synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 36–52.
[20] H. Dou, C. Chen, X. Hu, and S. Peng, "Asymmetric cyclegan for unpaired NIR-to-RGB face image translation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 1757–1761.
[21] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 687–704.
[22] H. Tang, D. Xu, H. Liu, and N. Sebe, "Dual generator generative adversarial networks for multi-domain image-to-image translation," in *Proc. ACCV*, 2018.
[23] H. Fu *et al.*, "Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2427–2436.
[24] H. Chang, J. Lu, F. Yu, and A. Finkelstein, "PairedCycleGAN: Asymmetric style transfer for applying and removing makeup," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 40–48.
[25] Y. Li *et al.*, "Asymmetric GAN for unpaired image-to-image translation," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5881–5896, Dec. 2019.
[26] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4500–4509.
[27] E. de Bézenac, I. Ayed, and P. Gallinari, "Optimal unsupervised domain translation," *ICLR*, 2019.
[28] Y. Yang *et al.*, "Video captioning by adversarial LSTM," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5600–5611, Nov. 2018.
[29] X. Xu *et al.*, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2400–2413, Jun. 2020.
[30] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 154–162.
[31] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "DeblurGAN: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8183–8192.
[32] R. Anirudh, J. J. Thiagarajan, B. Kailkhura, and T. Bremer, "An unsupervised approach to solving inverse problems using generative adversarial networks," 2018, *arXiv:1805.07281*.
[33] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.
[34] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[35] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 649–666.
[36] H. Tang, D. Xu, W. Wang, Y. Yan, and N. Sebe, "Dual generator generative adversarial networks for multi-domain image-to-image translation," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 3–21.
[37] N. Li *et al.*, "The synthesis of unpaired underwater images using a multistyle generative adversarial network," *IEEE Access*, vol. 6, pp. 54241–54257, 2018.
[38] Y.-F. Zhou *et al.*, "BranchGAN: Unsupervised mutual image-to-image transfer with a single encoder and dual decoders," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3136–3149, Dec. 2019.
[39] L. Chen, L. Wu, Z. Hu, and M. Wang, "Quality-aware unpaired image-to-image translation," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2664–2674, Oct. 2019.
[40] J. Wei, Y. Yang, X. Xu, X. Zhu, and H. T. Shen, "Universal weighting metric learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.
[41] H. T. Shen *et al.*, "Exploiting subspace relation in semantic labels for cross-modal hashing," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 10, pp. 3351–3365, Oct. 2021.
[42] M. Liu *et al.*, "STGAN: A unified selective transfer network for arbitrary image attribute editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3673–3682.
[43] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
[44] C. Shen *et al.*, "Training generative adversarial networks in one stage," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3350–3360.
[45] Y. Zou *et al.*, "PseudoSeg: Designing pseudo labels for semantic segmentation," in *Proc. Int. Conf. Learn. Representation*, 2021.
[46] Y.-H. Tsai, K. Sohn, S. Schulter, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1456–1465.
[47] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.
[48] F. Yu *et al.*, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2636–2645.
[49] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "APDrawingGAN: Generating artistic portrait drawings from face photos with hierarchical GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10743–10752.
[50] J.-Y. Zhu *et al.*, "Toward multimodal image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 465–476.
[51] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.
[52] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[53] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. Int. Conf. Pattern Recognit.*, 2010, pp. 2366–2369.
[54] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4401–4410.
[55] D. Yang, S. Hong, Y. Jang, T. Zhao, and H. Lee, "Diversity-sensitive conditional generative adversarial networks," in *Proc. Int. Conf. Learn. Representation*, 2019.
[56] R. Tyleček and R. Šára, "Spatial pattern templates for recognition of objects with regular structure," in *Proc. German Conf. Pattern Recognit.*, 2013, pp. 364–374.

Ziqiang Zheng received the B.Eng. degree in communication engineering from the Ocean University of China, Qingdao, China, in 2019. He is currently a Research Assistant with the Center for Future Media, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include multimedia content analysis and computer vision.

Yi Bin received the Ph.D. degree from the University of Electronic Science and Technology of China, Chengdu, China in 2020. He is currently with the University of Electronic Science and Technology of China. His research interests include multimedia analysis, vision understanding, and deep learning.

Xiaoou Lv received the Ph.D. degree in statistical machine learning from the Department of Statistical Science, University College London, London, United Kingdom, in 2020. His research interests include deep generative models, transfer learning, few-shot learning, and weakly supervised learning.

Yang Wu (Member, IEEE) received the B.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2010, respectively. He is currently an Expert Researcher with Tencent. From July 2019 to May 2021, he was a Program-Specific Senior Lecturer with the Department of Intelligence Science and Technology, Kyoto University, Kyoto, Japan. From December 2014 to June 2019, he was an Assistant Professor with the NAIST International Collaborative Laboratory for Robotics Vision, Nara Institute of Science and Technology, Ikoma, Japan. From 2011 to 2014, he was a Program-Specific Researcher with the Academic Center for Computing and Media Studies, Kyoto University. His research interests include the fields of computer vision, pattern recognition, and image/video search and retrieval.

Yang Yang (Senior Member, IEEE) received the bachelor's degree in computer science from Jilin University, Changchun, China, in 2006, the master's degree in computer science from Peking University, Beijing, China, in 2009, and the Ph.D. degree in computer science from The University of Queensland, Brisbane, QLD, Australia, in 2012. He is currently with the University of Electronic Science and Technology of China, Chengdu, China. His current research interests include multimedia content analysis, computer vision, and social media analytics.

Heng Tao Shen (Fellow, IEEE) received the B.Sc. (with first-class Hons.) and Ph.D. degrees from the Department of Computer Science, National University of Singapore, Singapore, in 2000 and 2004, respectively. He is currently the Dean of the School of Computer Science and Engineering, the Executive Dean of the AI Research Institute, University of Electronic Science and Technology of China, Chengdu, China. His research interests mainly include multimedia search, computer vision, artificial intelligence, and Big Data management. He is/was an Associate Editor for the *ACM Transactions of Data Science*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, and *Pattern Recognition.* He is a Fellow of ACM and OSA.