Perceptual Adversarial Networks for Image-to-Image Transformation

Chaoyue Wang, Chang Xu, Chaohui Wang, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— In this paper, we propose perceptual adversarial networks (PANs) for image-to-image transformations. Different from existing application driven algorithms, PAN provides a generic framework of learning to map from input images to desired images (Fig. 1), such as a rainy image to its de-rained counterpart, object edges to photos, and semantic labels to a scenes image. The proposed PAN consists of two feed-forward convolutional neural networks: the image transformation network *T* and the discriminative network *D*. Besides the generative adversarial loss widely used in GANs, we propose the perceptual adversarial loss, which undergoes an adversarial training process between the image transformation network *T* and the hidden layers of the discriminative network *D*. The hidden layers and the output of the discriminative network *D* are upgraded to constantly and automatically discover the discrepancy between the transformed image and the corresponding ground truth, while the image transformation network *T* is trained to minimize the discrepancy explored by the discriminative network *D*. Through integrating the generative adversarial loss and the perceptual adversarial loss, *D* and *T* can be trained alternately to solve image-to-image transformation tasks. Experiments evaluated on several image-to-image transformation tasks (e.g., image de-raining and image inpainting) demonstrate the effectiveness of the proposed PAN and its advantages over many existing works.**

***Index Terms*— Generative adversarial networks, image de-raining, image inpainting, image-to-image transformation.**

I. Introduction

IMAGE-TO-IMAGE transformations aim to transform an input image into the desired output image, and they exist in a number of applications about image processing, computer graphics, and computer vision. For example, generating high-quality images from corresponding degraded (e.g. simplified, corrupted or low-resolution) images, and transforming a color input image into its semantic or geometric representations. More examples include, but not limited to, image de-noising [1], image in-painting [2], image super-resolution [3], image colorization [4], image segmentation [5], *etc*.

In recent years, convolutional neural networks (CNNs) are trained in a supervised manner for various image-to-image transformation tasks [6]–[9]. They encode input image into hidden representation, which is then decoded to the output image. By penalizing the discrepancy between the output image and ground-truth image, optimal CNNs can be trained to discover the mapping from the input image to the transformed image of interest. These CNNs are developed with distinct motivations and differ in the loss function design.

One of the most straightforward approaches is to pixel-wisely evaluate output images [8]–[12], *e.g.*, least squares or least absolute losses to calculate the distance between the output and ground-truth images in the pixel space. Though pixel-wise evaluation can generate reasonable images, there are some unignorable defects associated with the outputs, such as image blur and image artifacts.

Besides pixel-wise losses, the generative adversarial losses were largely utilized in training image-to-image transformation models. GANs (and cGANs) [13], [14] perform an adversarial training process alternating between identifying and faking, and generative adversarial losses are formulated to evaluate the discrepancy between the generated distribution and the real-world distribution. Experimental results show that generative adversarial losses are beneficial for generating more realistic images. Therefore, there are many GANs (or cGANs) based works to solve image-to-image transformation tasks, resulting in sharper and more realistic transformed images [7], [15]. Meanwhile, some GANs variants [16]–[19] investigated cross-domain image translations and performed image translations in absence of paired examples. Although these unpaired works achieved reasonable results in some image-to-image translation tasks, they are inappropriate for some image-to-image problems. For example, in image in-painting tasks, it is difficult to define the domain and formulate the distribution of corrupted images. In addition, paired information within training data are beneficial for learning image transformations, but they cannot be utilized by unpaired translation methods.

Moreover, perceptual losses emerged as a novel measurement for evaluating the discrepancy between high-level perceptual features of the output and ground-truth images [20]–[22]. Hidden layers of a well-trained image classification network (*e.g.*, VGG-16 [23]) are usually employed to extract high-level features (*e.g.*, content or texture) of both

Manuscript received July 24, 2017; revised January 23, 2018 and April 4, 2018; accepted April 26, 2018. Date of publication May 14, 2018; date of current version May 24, 2018. This work was supported in part by the Australian Research Council Projects under Projects FL-170100117, DE-180101438, DP-180103424, and LP-150100671, and in part by SAP SE and CNRS under Grant INS2IJCJC-INVISANA. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Catarina Brites. *(Corresponding author: Chang Xu.)*

C. Wang is with the Centre for Artificial Intelligence, Faculty of Engineering and Information Technologies, School of Software, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: chaoyue.wang@student.uts.edu.au).

C. Xu and D. Tao are with the UBTECH Sydney Artificial Intelligence Centre, Faculty of Engineering and Information Technologies, School of Information Technologies, The University of Sydney, Darlington, NSW 2008, Australia (e-mail: c.xu@sydney.edu.au; dacheng.tao@sydney.edu.au).

C. Wang is with the LIGM Lab, UMR 8049 CNRS·ENPC·ESIEE·UPEM, Marne-la-Vallée, Université Paris-Est, 77420 Champs-sur-Marne, France (e-mail: chaohui.wang@u-pem.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2836316

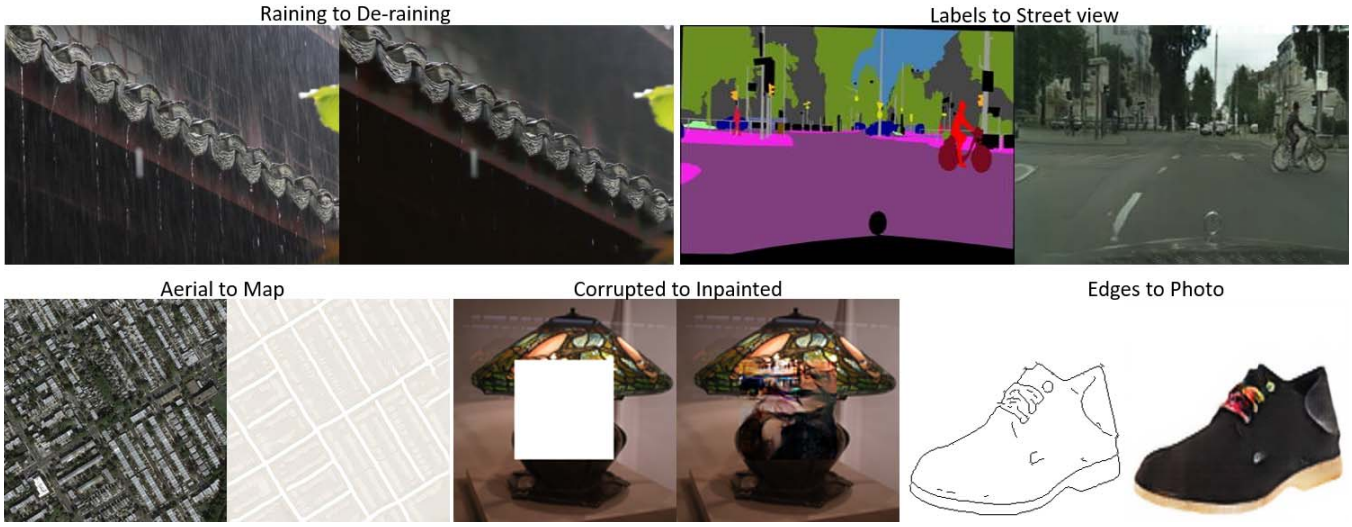

Fig. 1. Image-to-image transformation tasks. Many tasks in image processing, computer graphics, and computer vision can be regarded as image-to-image transformation tasks, where a model is designed to transform an input image into the required output image. We proposed Perceptual Adversarial Networks (PAN) to solve the image-to-image transformation between paired images. For each pair of the images we demonstrated, the left one is the input image, and the right one is the transformed result of the proposed PAN.

output images and ground-truth images. It is then expected to encourage the output image to have the similar high-level feature with that of the ground-truth image. Recently, perceptual losses were introduced in aforementioned GANs-based image-to-image transformation frameworks for suppressing artifacts [24] and improving perceptual quality [25], [26] of the output images. Though integrating perceptual losses into GANs has produced impressive image-to-image transformation results, existing works are used to depend on external well-trained image classification network (*e.g.*VGG-Net) out of GANs, but ignored the fact that GANs, especially the discriminative network, also has the capability and demand of perceiving the content of images and the difference between images. Moreover, since these external networks are trained on specific classification datasets (*e.g.*, ImageNet), they mainly focus on features that contribute to the classification and may perform inferior in some image transformation tasks (*e.g.*, transfer aerial images to maps). Meanwhile, since specific hidden layers of pre-trained networks are employed, it is difficult to explore the difference between generated images and ground-truth images from more points of view.

In this paper, we proposed the perceptual adversarial networks (PAN) for image-to-image transformation tasks. Inspired by GANs, PAN is composed of an image transformation network T and a discriminative network D. Both generative adversarial loss and perceptual adversarial loss are employed. Firstly, similar with GANs, the generative adversarial loss is utilized to measure the distribution of generated images, *i.e.*, penalizing generated images to lie in the desired target domain, which usually contributes to producing more visually realistic images. Meanwhile, to comprehensively evaluate transformed images, we devised the perceptual adversarial loss to form dynamic measurements based on the hidden layers of the discriminative network D. Specifically, given hidden layers of the network D, the network T is trained to generate the output image that has the same high-level features with that of the corresponding ground-truth. If the difference between images measured on existing hidden layers of the discriminator is smaller, these hidden layers will be updated to discover the discrepancy between images from a new point of view. Different from the pixel-wise loss and conventional perceptual loss, our perceptual adversarial loss undergoes an adversarial training process, and aims to discover and decrease the discrepancy under constantly explored dynamic measurements.

In summary, our paper makes the following contributions:

- We proposed the perceptual adversarial loss, which utilizes the hidden layers of the discriminative network to evaluate the discrepancy between the output and ground-truth images through an adversarial training process.
- Through combining the perceptual adversarial loss and the generative adversarial loss, we presented the PAN for solving image-to-image transformation tasks.
- We evaluated the performance of the PAN on several image-to-image transformation tasks (Fig. 1). Experimental results show that the proposed PAN has a great capability of accomplishing image-to-image transformations.

The rest of the paper is organized as follows: after a brief summary of previous related works in section II, we illustrate the proposed PAN together with its training losses in section III. Then we exhibit the experimental validation of the whole method in section IV. Finally, we conclude this paper with some future directions in section V.

II. Background

In this section, we first introduce some representative image-to-image transformation methods based on feed-forward CNNs, and then summarize related works on GANs and perceptual losses.

A. Image-to-Image Transformation With Feed-Forward CNNs

Recent years have witnessed a variety of feed-forward CNNs developed for image-to-image transformation tasks. These feed-forward CNNs can be easily trained using the back-propagation algorithm [27], and the transformed images are generated by forwardly passing the input image through the well-trained CNNs in the test stage.

Individual pixel-wise loss or pixel-wise loss accompanied with other losses are employed in a number of image-to-image transformations. Image super-resolution tasks estimate a high-resolution image from its low-resolution counterpart [8], [20], [25]. Image de-raining (or de-snowing) methods attempt to remove the rain (or snow) strikes in the pictures brought by the uncontrollable weather conditions [6], [24], [28]. Given a damaged image, image inpainting aims to recover the missing part of the input image [7], [29], [30]. Image semantic segmentation methods produce dense scene labels based on a single input image [31]–[33]. Given an input object image, some feed-forward CNNs were trained to synthesize the image of the same object from a different viewpoint [34], [35]. More image-to-image transformation tasks based on feed-forward CNNs, include, but not limited to, image colorization [10], depth estimations [33], [36], *etc*.

B. GANs-Based Works

Generative adversarial networks (GANs) [13] provide an important approach for learning a generative model which generates samples from the real-world data distribution. GANs consist of a generative network and a discriminative network. Through playing a minimax game between these two networks, GANs are trained to generate more and more 'realistic samples. Since the great performance on learning real-world distributions, there have emerged a large number of GANs-based works. Some of these GANs-based works are committed to training a better generative model, such as InfoGAN [37], Energy-based GAN [38], WGAN(-GP) [39], [40], Progressive GAN [41], E-GAN [42] and SN-GAN [43]. There are also some works integrating the GANs into their models to improve the performance of classical tasks. For example, the PGAN [44] is proposed for small object detection. Specifically, Li *et al.* [44] devised a novel perceptual discriminator network, which contains an adversarial branch and a perception branch. The adversarial branch utilizes the adversarial loss to distinguish representations of real and synthesized objectives; the perception branch (or loss) employs a classification loss L_{cls} and a bounding-box regression loss L_{loc} to encourage the synthesized 'super-resolved' objectives representation to retain the same perception information as the input small objectives representation.

In addition, these kind of works include, but not limited to, the PGN [45] for video prediction, the SRGAN [25] for super-resolution, the ID-CGAN for image de-raining [24], the iGAN [46] for interactive application, the IAN [47] for photo modification, and the Context-Encoder for image inpainting [7]. Most recently, Isola *et al.* [15] proposed the pix2pix-cGANs to perform several image-to-image transformation tasks (also known as image-to-image translations in their work), such as translating semantic labels into the street scene, object edges into pictures, aerial photos into maps, *etc*.

Moreover, some GANs variants [16]–[18] investigated cross-domain image translations through exploring the cyclic mapping (or primal-dual) relation between different image domains. Specifically, a primal GAN aims to explore the mapping relations from source images to target images, while a dual (or inverse) GAN performs the invert task. These two GANs form a closed loop and allow images from either domain to be translated and then reconstructed. Through combining the GAN loss and cycle consistency loss (or recovery loss), these works can be used for performing image translation tasks in absence of paired examples. However, if paired training data are available in some applications, Zhu *et al.* [16], Yi *et al.* [17], and Kim *et al.* [18] neglect paired information between data often have inferior performance to that of paired methods [15]. Thus, at this stage, it is still important to study paired training, especially for performance-driven situations and applications, such as high-resolution image synthesis [48], photo-realistic image synthesis [25], real-world image inpainting [49], *etc*.

C. Perceptual Loss

Recently, some theoretical analysis and experimental results suggested that the high-level features extracted from a well-trained image classification network have the capability to capture the perceptual information from real-world images [20], [50]. Specifically, representations extracted from hidden layers of well-trained image classification network are beneficial to interpret the semantics of input images, and image style distribution can be captured by the Gram matrix of hidden representations. Hence, high-level features extracted from hidden layers of a well-trained classifier are often introduced in image generation models. Dosovitskiy and Brox [21] took Euclidean distances between high-level features of images as the deep perceptual similarity metrics to improve the performance of image generation. Johnson *et al.* [20], Bruna *et al.* [22], and Ledig *et al.* [25] used features extracted from a well-trained VGG network to improve the performance of single image super-resolution task. In addition, there are works applying high-level features in image style-transfer [20], [50], image de-raining [24] and image view synthesis [51] tasks.

III. Methods

In this section, we introduce the proposed Perceptual Adversarial Networks (PAN) for image-to-image transformation tasks. Firstly, we explain the generative and perceptual adversarial losses, respectively. Then, we give the whole framework of the proposed PAN. Finally, we illustrate the details of the training procedure and network architectures.

A. Generative Adversarial Loss

We begin with the generative adversarial loss in vanilla GANs. A generative network G is trained to map samples from noise distribution p_z to real-world data distribution p_{data} through playing a minimax game with a discriminative network D. In the training procedure, the discriminative network

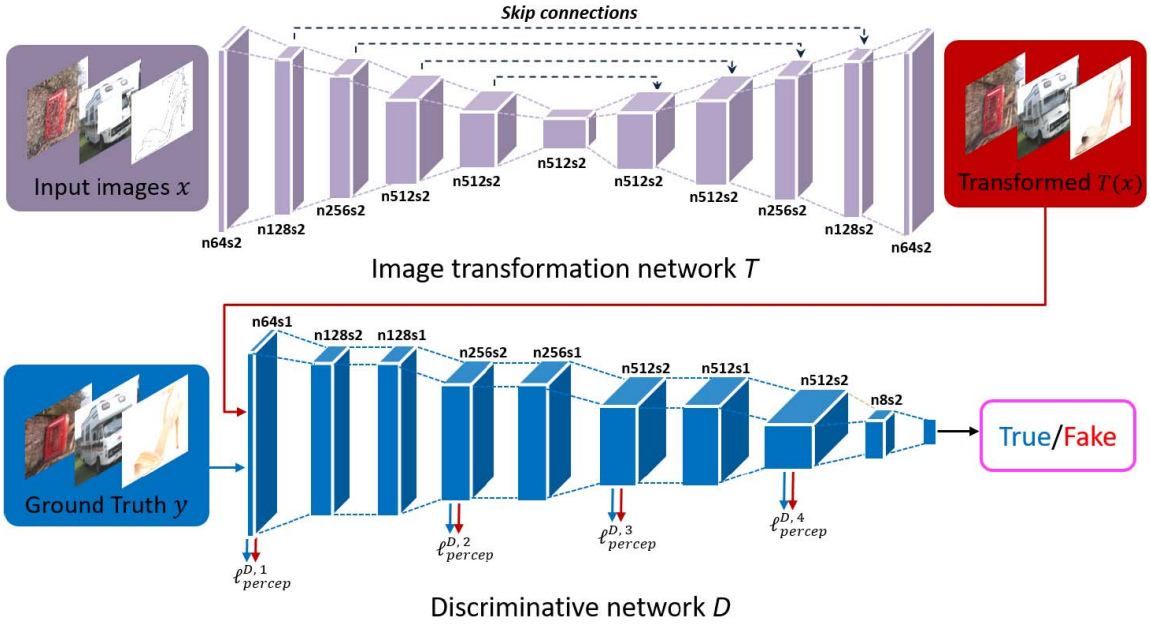

Fig. 2. PAN framework. PAN consists of an image transformation network T and a discriminative network D. The image transformation network T is trained to synthesize the transformed images given the input images. It is composed of a stack of Convolution-BatchNorm-LeakyReLU encoding layers and Deconvolution-BatchNorm-ReLU decoding layers, and the skip-connections are used between mirrored layers. The discriminative network D is also a CNN that consists of Convolution-BatchNorm-LeakyReLU layers. Hidden layers of the network D are utilized to evaluate the perceptual adversarial loss, and the output of the network D is used to distinguish transformed images from real-world images.

D aims to distinguish the real samples $y \sim p_{\text{data}}$ from the generated samples $G(z)$. In contrary, the generative network G tries to confuse the discriminative network D by generating increasingly realistic samples. This minimax game can be formulated as:

$$\min_G \max_D \mathbb{E}_{y\sim p_{\text{data}}}[\log D(y)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \quad (1)$$

Nowadays, GANs-based models have shown the strong capability of learning generative models, especially for image generation [37], [39], [45]. We, therefore, adopt the GANs learning strategy to solve image-to-image transformation tasks as well. As shown in Fig. 2, the image transformation network T is used to generate transformed image $T(x)$ given the input image $x \in \mathcal{X}$. Meanwhile, each input image x has a corresponding ground-truth image y. We suppose that all target image $y \in \mathcal{Y}$ obey the distribution p_{real}, and the transformed image $T(x)$ is encouraged to have the same distribution with that of targets image y, *i.e.*, $T(x) \sim p_{\text{real}}$. To achieve the generative adversarial learning strategy, a discriminative network D is additionally introduced, and the generative adversarial loss can be written as:

$$\min_T \max_D \mathcal{V}_{D,T} = \mathbb{E}_{y\in\mathcal{Y}}[\log D(y)] + \mathbb{E}_{x\in\mathcal{X}}[\log(1 - D(T(x)))] \quad (2)$$

The generative adversarial loss acts as a statistical measurement to penalize the discrepancy between the distributions of transformed images and the ground-truth images.

B. Perceptual Adversarial Loss

Different from vanilla GANs that randomly generate samples from the data distribution p_{data}, our goal is to infer the transformed image according to the input images. Therefore, it is a further step of GANs to explore the mapping from the input image to its ground truth.

As mentioned in Sections I and II, pixel-wise losses and perceptual losses are widely used in existing works for generating images towards the ground truth. The pixel-wise losses penalize the discrepancy occurred in the pixel space, but often produce blurry results [7], [9]. The perceptual losses explore the discrepancy between high-dimensional representations of images extracted from a well-trained classifier, *e.g.*, the VGG net trained on the ImageNet dataset [23]. Although hidden layers of well-trained classifier have been experimentally validated to map the image from pixel space to high-level feature spaces, how to extract the effective features for image-to-image transformation tasks from hidden layers has not been thoroughly discussed.

Here, we employ hidden layers of the discriminative network D to evaluate the perceptual adversarial loss between transformed images and ground-truth images. In our experiments, given the training sample $\{(x_i, y_i) \in (\mathcal{X} \times \mathcal{Y})\}_{i=1}^N$, the least absolute loss is employed to calculate the discrepancy of the high-dimensional representations on the hidden layers of the network D, *e.g.*,

$$\ell^{D,j}_{percep} = \frac{1}{N}\sum_{i=1}^{N} ||d_j(y_i) - d_j(T(x_i))|| \quad (3)$$

where $d_j(\cdot)$ is the image representation on the j^{th} hidden layer of the discriminative network D, and $\ell^{D,j}_{percep}$ calculates the discrepancy measured by the j^{th} hidden layer of D.

Similar to what has been done with the Energy-Based GAN [38], we use two different losses, one ($\mathcal{L}_T$) to train

the image transformation network T, and the other ($\mathcal{L}_D$) to train hidden layers of the discriminative network D. Therefore, the image transformation network T and hidden layers of the discriminative network D play a non-zero-sum game and form the perceptual adversarial loss. Formally, the perceptual adversarial loss $\mathcal{L}_T$ for the image transformation network T can be written as:

$$\mathcal{L}_T = \sum_{j=1}^{F} \lambda_j \ell_{percep}^{D,j} \tag{4}$$

and, given a positive margin m, the loss $\mathcal{L}_D$ for hidden layers of the discriminative network D is defined as:

$$\mathcal{L}_D = [m - \mathcal{L}_T]^+ = \left[m - \sum_{j=1}^{F} \lambda_j \ell_{percep}^{D,j}\right]^+ \tag{5}$$

where $[\cdot]^+ = \max(0, \cdot)$, $\{\lambda_j\}_{j=1}^F$ are hyper-parameters balancing the influence of F different hidden layers.

By minimizing the perceptual adversarial loss function $\mathcal{L}_T$ with respect to parameters of T, we encourage the network T to generate image $T(x)$ that has similar high-level features with its ground-truth y on the hidden layers. If the weighted sum of discrepancy between transformed images and ground-truth images on different hidden layers is less than the positive margin m, the loss function $\mathcal{L}_D$ will upgrade the discriminative network D for some new latent feature spaces, which preserve the discrepancy between the transformed images and their ground-truth. Therefore, based on the perceptual adversarial loss, the discrepancy between the transformed and ground-truth images can be constantly explored and exploited.

Compared to our perceptual adversarial loss which measures the difference between the transformed image and ground-truth image in hidden layers of the discriminator, the conditional GAN loss indicates whether the transformed image forms the appropriate image pair with the input image, and can also explore supervised information of paired images during the training process. However, they utilize different methods to minimize high-level feature differences explored by the discriminator. The perceptual adversarial loss directly penalizes the high-level representations of transformed images and ground-truth images to be as same as possible. In contrast, the conditional GAN loss aims to model the mapping relation from the input x to its output y_{real} and encourages the generated image pairs $(x, T(x))$ obeying the same conditional distribution $P_{\text{real}}(y|x)$. Compared to conditional GAN loss that indirectly guides the generated images $T(x)$ sharing the same features with corresponding ground-truth y_{real}, our perceptual adversarial loss directly measures and minimizes differences between generated images and ground-truth images from different perspectives.

C. The Perceptual Adversarial Networks

Based on the aforementioned generative adversarial loss (Eq. 2) and perceptual adversarial loss (Eq. 4 and Eq. 5), we develop the PAN framework, which consists of an image transformation network T and a discriminative network D. These two networks are trained alternately to perform an adversarial learning process, the loss functions of image transformation network $\mathcal{J}_T$ and discriminative network $\mathcal{J}_D$ are formally defined as:

$$\begin{aligned} \mathcal{J}_T &= \theta \mathcal{V}_{D,T} + \mathcal{L}_T \\ \mathcal{J}_D &= -\theta \mathcal{V}_{D,T} + \mathcal{L}_D \\ &= -\theta \mathcal{V}_{D,T} + [m - \mathcal{L}_T]^+ \end{aligned} \tag{6}$$

where θ is the hyper-parameter balance the influence of generative adversarial loss and perceptual adversarial loss. When $\mathcal{L}_T < m$, minimizing $\mathcal{J}_D$ with respect to the parameters of D is consistent with maximizing $\mathcal{J}_T$. Otherwise, when $\mathcal{L}_T \geq m$, the second term of $\mathcal{J}_D$ will have zero gradients, because of the positive margin m. In general, the discriminative network D aims to distinguish transformed image $T(x)$ from ground-truth image y from both the statical (the first term of $\mathcal{J}_D$) and dynamic perceptual (the second term of $\mathcal{J}_D$) aspects. On the other hand, the image transformation network T is trained to generate increasingly better images by reducing the discrepancy between the output and ground-truth images.

D. Network Architectures

Fig. 2 illustrates the framework of the proposed PAN, which is composed of two CNNs, *i.e.*, the image transformation network T and the discriminative network D.

1) Image Transformation Network T: The image transformation network T is designed to generate the transformed image given the input image. Following the network architectures in [15] and [52], the network T firstly encodes the input image into high-dimensional representation using a stack of Convolution-BatchNorm-LeakyReLU layers, and then, the output image can be decoded by the following Deconvolution-BatchNorm-ReLU layers.[1] Note that the output layer of the network T does not use batchnorm and replaces the ReLU with Tanh activation. Moreover, the skip-connections are used to connect mirrored layers in the encoder and decoder stacks. More details of the transformation network T are listed in Table II. The same architecture of the network T is used for all experiments in this paper, except there is an additional explanation.[2]

2) Discriminative Network D: In the proposed PAN framework, the discriminative network D is introduced to compute the discrepancy between the transformed images and the ground-truth images. Specifically, given an input image, the discriminative network D extracts high-level features using a series of Convolution-BatchNorm-LeakyReLU layers. The 1^{st}, 4^{th}, 6^{th}, and 8^{th} layers are utilized to measure the perceptual adversarial loss for every pair of transformed image and its corresponding ground-truth in the training data. Finally, the last convolution layer is flattened and then fed into a single sigmoid output. The output of the discriminative network D estimates the probability that the input image comes from the real-world dataset rather than from the image transformation network T. The same discriminative network D is applied for

[1] The deconvolution layer utilized in our framework is the transposed convolution layer used in [53] and [54].

[2] In the analysis of the loss functions and the image inpainting task, different architectures of the network T were used.

TABLE I
THE ARCHITECTURE OF THE DISCRIMINATIVE NETWORK

Discriminative network D	
Input: Image	
[layer 1]	Conv. (3, 3, 64), stride=1; *LReLU*;
	(Perceptual adversarial loss: $\ell_{percep}^{D,1}$)
[layer 2]	Conv. (3, 3, 128), stride=2; Batchnorm; *LReLU*;
[layer 3]	Conv. (3, 3, 128), stride=1; Batchnorm; *LReLU*;
[layer 4]	Conv. (3, 3, 256), stride=2; Batchnorm; *LReLU*;
	(Perceptual adversarial loss: $\ell_{percep}^{D,2}$)
[layer 5]	Conv. (3, 3, 256), stride=1; Batchnorm; *LReLU*;
[layer 6]	Conv. (3, 3, 512), stride=2; Batchnorm; *LReLU*;
	(Perceptual adversarial loss: $\ell_{percep}^{D,3}$)
[layer 7]	Conv. (3, 3, 512), stride=1; Batchnorm; *LReLU*;
[layer 8]	Conv. (3, 3, 512), stride=2; Batchnorm; *LReLU*;
	(Perceptual adversarial loss: $\ell_{percep}^{D,4}$)
[layer 9]	Conv. (3, 3, 8), stride=2; *LReLU*;
[layer 10]	Fully connected (1); *Sigmoid*;
Output: Real or Fake (Probability)	

TABLE II
THE ARCHITECTURE OF THE IMAGE TRANSFORMATION NETWORK

Image transformation network T	
Input: Image	
[layer 1]	Conv. (3, 3, 64), stride=2; *LReLU*;
[layer 2]	Conv. (3, 3, 128), stride=2; Batchnorm;
[layer 3]	*LReLU*; Conv. (3, 3, 256), stride=2; Batchnorm;
[layer 4]	*LReLU*; Conv. (3, 3, 512), stride=2; Batchnorm;
[layer 5]	*LReLU*; Conv. (3, 3, 512), stride=2; Batchnorm;
[layer 6]	*LReLU*; Conv. (3, 3, 512), stride=2; Batchnorm; *LReLU*;
[layer 7]	DeConv. (4, 4, 512), stride=2; Batchnorm;
	Concatenate Layer(Layer 7, Layer 5); *ReLU*;
[layer 8]	DeConv. (4, 4, 256), stride=2; Batchnorm;
	Concatenate Layer(Layer 8, Layer 4); *ReLU*;
[layer 9]	DeConv. (4, 4, 128), stride=2; Batchnorm;
	Concatenate Layer(Layer 9, Layer 3); *ReLU*;
[layer 10]	DeConv. (4, 4, 64), stride=2; Batchnorm;
	Concatenate Layer(Layer 10, Layer 2); *ReLU*;
[layer 11]	DeConv. (4, 4, 64), stride=2; Batchnorm; *ReLU*;
[layer 12]	DeConv. (4, 4, 3), stride=2; *Tanh*;
Output: Transformed image	

all tasks demonstrated in this paper, and details of the network D are shown in Table I.

IV. EXPERIMENTS

In this section, we evaluate the performance of the proposed PAN on several image-to-image transformation tasks, which are popular in fields of image processing (*e.g.*, image de-raining), computer vision (*e.g.*, semantic segmentation) and computer graphics (*e.g.*, image generation).

A. Experimental Setting Up

For fair comparisons, we adopted the same settings with existing works, and reported experimental results using several evaluation metrics. These tasks and data settings include:

- *Single image de-raining*, on the dataset provided by ID-CGAN [24].
- *Image Inpainting*, on a subset of ILSVRC'12 (same as context-encoder [7]).
- *Semantic labels↔images*, on the Cityscapes dataset [55] (same as pix2pix [15]).
- *Edges→images*, on the dataset created by pix2pix [15]. The original data is from [46] and [56], and the HED edge detector [57] was used to extract edges.
- *Aerial→map*, on the dataset from pix2pix [15].

Furthermore, all experiments were trained on Nvidia Titan-X GPUs using Theano [58]. Given the generative and perceptual adversarial losses, we alternately updated the image transformation network T and the discriminative network D. Specifically, Adam solver [59] with a learning rate of 0.0002 and a first momentum of 0.5 was used in network training. After one update of the discriminative network D, the image transformation T will be updated three times. Hyper-parameters $\theta = 1$, $\lambda_1 = 5$, $\lambda_2 = 1.5$, $\lambda_3 = 1.5$, $\lambda_4 = 1$, and batch size of 4 were used for all tasks. Since the dataset sizes for different tasks are changed largely, the training epochs of different tasks were set accordingly. Overall, the number of training iterations was around 100k.

B. Evaluation Metrics

To illustrate the performance of image-to-image transformation tasks, we conducted qualitative and quantitative experiments to evaluate the performance of the transformed images. For the qualitative experiments, we directly presented the input and transformed images. Meanwhile, we used quantitative measures to evaluate the performance over the test sets, such as Peak Signal to Noise Ratio (PSNR), Structural Similarity Index (SSIM) [60], Universal Quality Index (UQI) [61] and Visual Information Fidelity (VIF) [62]. f training iterations were around 100k.

C. Analysis of the Loss Functions

As discussed in Sections I and II, the design of loss function will largely influence the performance of image-to-image transformation. Firstly, the pixel-wise loss (using least squares loss) is widely used in various image-to-image transformation works [63], [64]. Then, the joint loss integrating pixel-wise loss and conditional generative adversarial loss is proposed to synthesize more realistic transformed images [7], [15]. Most recently, through introducing the perceptual loss, *i.e.*, penalizing the discrepancy between high-level features that extracted by a well-trained classifier, the performance of some image-to-image transformation tasks are further

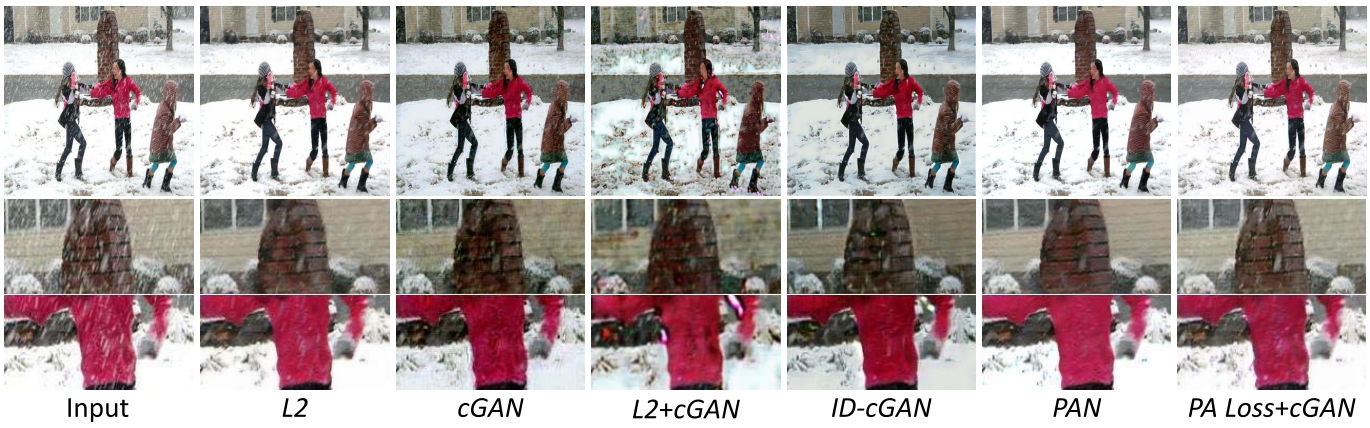

Fig. 3. Comparison of snow-streak removal using different losses functions. Given the same input image (leftmost), each column shows results trained under different losses. The loss function of ID-CGAN [24] combined the pixel-wise loss (least squares loss), cGANs loss and perceptual loss, *i.e.*, L2+cGAN+perceptual. For better visual comparison, zoomed versions of the specific regions-of-interest are demonstrated below the test images.

enhanced [20], [24], [25]. Different from these existing methods, the proposed PAN loss integrates the generative adversarial loss and the perceptual adversarial loss to train image-to-image transformation networks. Here, we compare the performance of the proposed perceptual adversarial loss with those of existing losses. For a fair comparison, we adopted the same image transformation network and data settings from ID-CGAN [24], and used the combination of different losses to perform the image de-raining (de-snowing) task. The quantitative results over the synthetic test set were shown in Table III, while the qualitative results on the real-world images were shown in Fig. 3. From both quantitative and qualitative comparisons, we find that only using the pixel-wise loss (least squares loss) achieved the worst result, and there are many snow-streaks in the transformed images (Fig. 3). Through introducing the cGANs loss, the de-snowing performance was indeed improved, but artifacts can be observed (Fig. 3) and the PSNR performance dropped (Table III). Combining the pixel-wise, cGAN and perceptual loss (VGG-16 [23]) together, *i.e.*, using the loss function of ID-CGAN [24], the quality of transformed images has been further improved on both observations and quantitative measurements. However, from Fig. 3, we observe that the transformed images have some color distortion compared to the input images. The proposed PAN loss (i.e., combining the perceptual adversarial loss and original GAN loss) not only removed most streaks without color distortion, but also achieved much better performance on quantitative measurements. Moreover, we evaluated the performance of combining conditional GAN loss and the perceptual adversarial loss. Comparing with using the cGAN loss independently, introducing the perceptual adversarial loss largely improves the model performance. Yet, comparing with the PAN loss, replacing the original GAN loss with its conditional version does not make a further improvement in both quantitative and qualitative comparisons.

Though variables of the discriminator network are optimized in iterations, the capability of hidden layers is constrained by network architecture. Therefore, in the proposed PAN, we selected four hidden layers of the discriminative network D to calculate the perceptual adversarial loss. We next proceed

TABLE III
DE-RAINING

	PSNR(dB)	SSIM	UQI	VIF
L2	22.77	0.7959	0.6261	0.3570
cGAN	21.87	0.7306	0.5810	0.3173
L2+cGAN	22.19	0.8083	0.6278	0.3640
ID-CGAN	22.91	0.8198	0.6473	0.3885
PAN	**23.35**	**0.8303**	**0.6644**	**0.4050**
PA Loss+cGAN	23.22	0.8078	0.6375	0.3904

to analyze the property of these hidden layers. Specifically, we trained four configurations of the PAN to perform the task of transforming the semantic labels to the cityscapes images. For each configuration, we set one hyper-parameter λ_i as 1, and the others $\{\lambda_1, \ldots, \lambda_{i-1}, \lambda_{i+1}, \ldots\}$ as 0, *i.e.*, we used only one hidden layer to evaluate the perceptual adversarial loss in each configuration. As shown in Fig. 4, the lower layers (*e.g.*, $\ell^{D,1}_{percep}$, $\ell^{D,2}_{percep}$) pay more attention to the patch-to-patch transformation and the color transformation, but the transformed images are blurry and lack of fine details. On the other hand, higher layers (*e.g.*, $\ell^{D,1}_{percep}$, $\ell^{D,2}_{percep}$) capture more high-frequency information, but lose the color information. Therefore, by integrating different properties of these hidden layers, the proposed PAN can be expected to achieve better performance, and the final results of this task are shown in Fig. 8 and Table V.

In our work, the balance between the perceptual adversarial loss and GAN loss is controlled by the hyper-parameters θ. In the task of transforming labels to facades, we vary the value of θ to test its influence on the proposed PAN. Qualitative samples are reported in Fig 5. As shown in Fig. 5, only using the perceptual adversarial loss (*i.e.*, $\theta = 0$) has already had the capability of synthesizing visually reasonable images from the input labels. Given the advantage of the GAN loss to promote more realistic images, the transformation performance gets better with increasing θ. However, with the continuous increasing of θ, the role of perceptual adversarial loss will

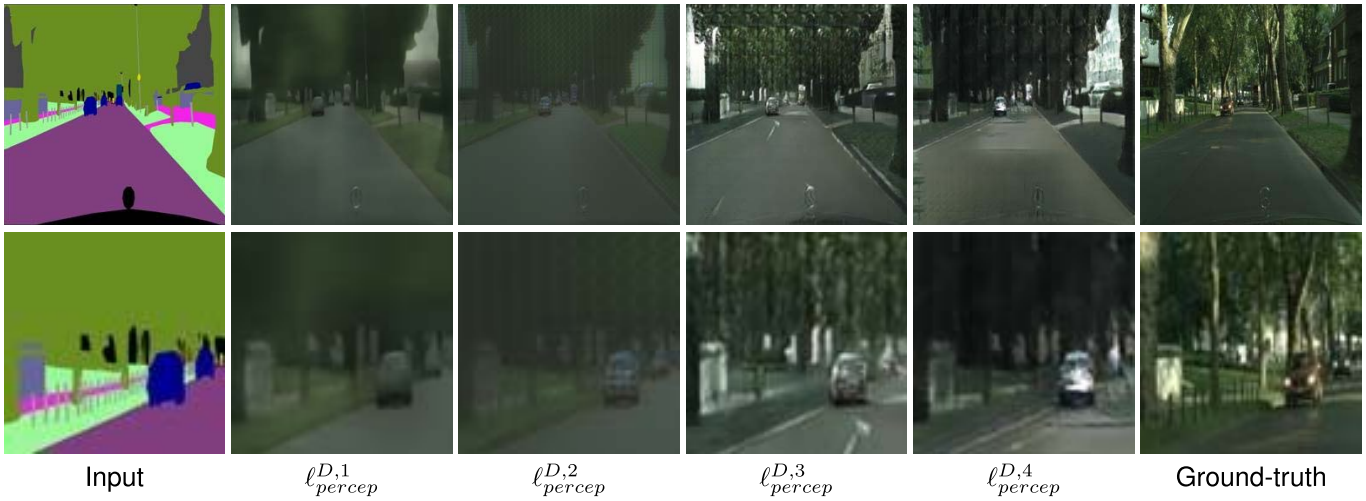

Fig. 4. Transforming the semantic labels to cityscapes images use the perceptual adversarial loss. Within the perceptual adversarial loss, a different hidden layer is utilized for each experiment. For better visual comparison, zoomed versions of the specific regions-of-interest are demonstrated below the test images. For higher layers, the transformed images look sharper, but less color information is preserved.

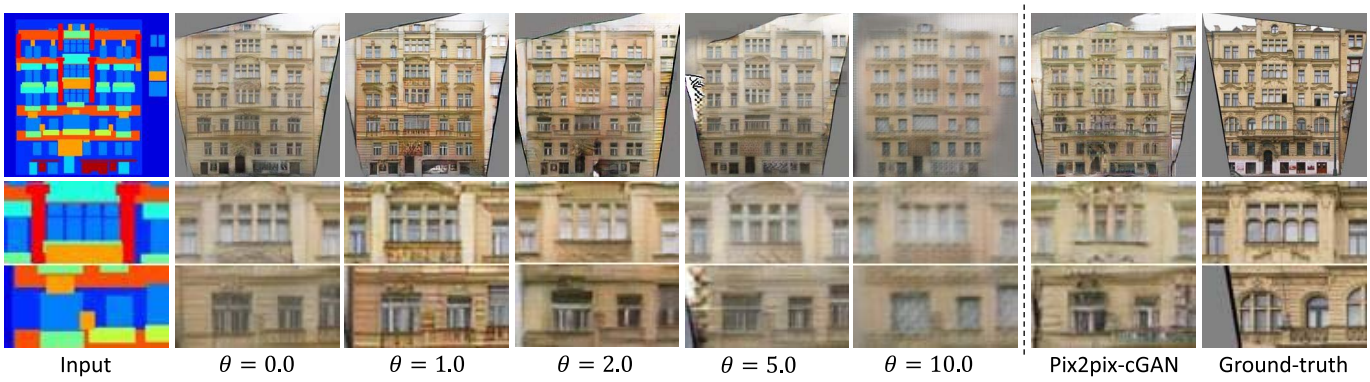

Fig. 5. Comparison of transforming the semantic labels to facades images by controlling the hyper-parameter θ. Given the same input image (leftmost), each column shows results trained under different θ. For better visual comparison, zoomed versions of the specific regions-of-interest are demonstrated below the test images.

be weakened and the model performance drops, *e.g.*, visual artifacts are observed in certain images.

D. Comparing With Existing Works

In this subsection, we compared the performance of the proposed PAN with those of existing algorithms for image-to-image transformation tasks.

1) Context-Encoder: Context-Encoder (CE) [7] trained CNNs for the single image inpainting task. Given corrupted images as input, image inpainting can be formulated as an image-to-image transformation task. Pixel-wise loss (least squares loss) and the generative adversarial loss were combined in the Context-Encoder to explore the relationship between the input surroundings and its central missed region.

To compare with the Context-Encoder, we applied PAN to inpaint images whose central regions were missed. As illustrated in Section IV-A, 100k images were randomly selected from the ILSVRC'12 dataset to train both Context-Encoder and PAN, and 50k images from the ILSVRC'12 validation set were used for test purpose. Moreover, since the image inpainting models are asked to generate the missing region of the input image instead of the whole image, we employ the image transformation network architecture from [7].

TABLE IV
IN-PAINTING

	PSNR(dB)	SSIM	UQI	VIF
Context-Encoder	21.74	0.8242	0.7828	0.5818
PAN	**21.85**	**0.8307**	**0.7956**	**0.6104**

In Fig. 7, we reported some example results in the test set. For each input image, the missing part is mixed by the foreground objects and backgrounds. From the inpainted results, we find the proposed PAN performed better on understanding the surroundings and estimating the missing part with semantic contents. However, the context-encoder tended to use the nearest region (usually the background) to inpaint the missing part. PAN can synthesize more details in the missing parts. Last but not the least, in Table IV, we reported the quantitative results calculated over all 50k test images,

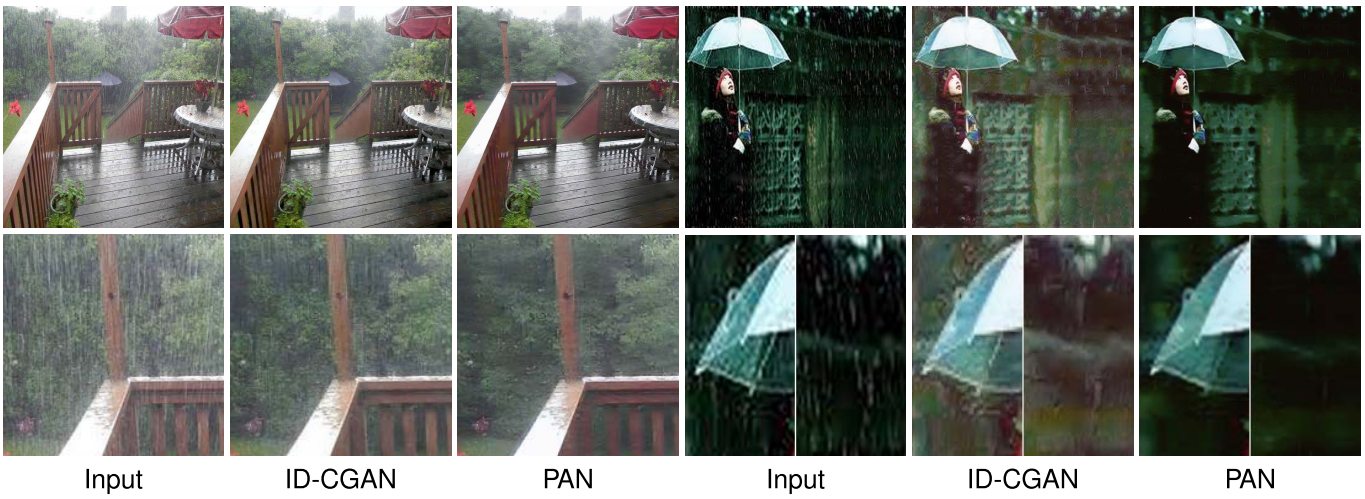

Fig. 6. Comparison of rain-streak removal using the ID-CGAN with the proposed PAN on real-world rainy images. For better visual comparison, zoomed versions of the specific regions-of-interest are demonstrated below the test images.

Fig. 7. Comparison of image in-painting results using the Context-Encoder(CE) with the proposed PAN. Given the central region missed input image (leftmost), the in-painted images and the ground-truth are listed on its rightside.

which also demonstrated that the proposed PAN achieves better performance.

2) ID-CGAN: Image de-raining task aims to remove rain streaks in a given rainy image. Considering the unpredictable weather conditions, the single image de-raining (de-snowing) is a challenge image-to-image transformation task. Most recently, the Image De-raining Conditional Generative Adversarial Networks (ID-CGAN) was proposed to tackle the image de-raining problem. Through combining the pixel-wise (least squares loss), conditional generative adversarial, and perceptual losses (VGG-16), ID-CGAN achieved the state-of-the-art performance on single image de-raining.

We attempted to solve image de-raining by the proposed PAN using the same setting with that of ID-CGAN. Since there is a lack of large-scale datasets consisting of paired rainy and de-rained images, we resort to synthesize the training set [24] of 700 images. Zhang *et al.* [24] provided 100 synthetic images and 50 real-world rainy images for the test. Since the ground-truth is available for synthetic test images, we calculated and reported the quantitative results in Table III. Moreover, we test both ID-CGAN and PAN on real-world rainy images, and the results were shown in Fig. 6. For better visual comparison, we zoomed up the specific regions-of-interest below the test images.

From Fig. 6, we found both ID-CGAN and PAN achieved great performance on single image de-raining. However, by observing the zoomed region, the PAN removed more rain-strikes with less color distortion. Additionally, as shown in Table III, for synthetic test images, the de-rained results of PAN are much more similar with the corresponding ground-truth than that of ID-CGAN. Dealing with the uncontrollable weather condition, why the proposed PAN can achieve better results? One possible reason is that ID-CGAN utilized the well-trained classifier to extract the high-level features of the output and ground-truth images, and penalize the discrepancy between them (*i.e.*, the perceptual loss). The high-level features extracted by the well-trained classifier usually focus on the content information, and may hard to capture other image information, such as color information. Yet, the proposed PAN used the perceptual adversarial loss, which aims to continually and automatically measure the discrepancy between the output and ground-truth images. The different training strategy of PAN may help the model to learn a better mapping from the input to output images, and resulting in better performance.

3) Pix2pix-cGAN: Isola *et al.* [15] utilized cGANs as a general-purpose solution to image-to-image translation (transformation) tasks. In their work, the pixel-wise loss (least absolute loss) and Patch-cGANs loss are employed to solve

Fig. 8. Comparison of transforming the semantic labels to cityscapes images using the pix2pix-cGAN with the proposed PAN. Given the semantic labels (leftmost), the transformed cityscapes images and the ground-truth are listed on the rightside.

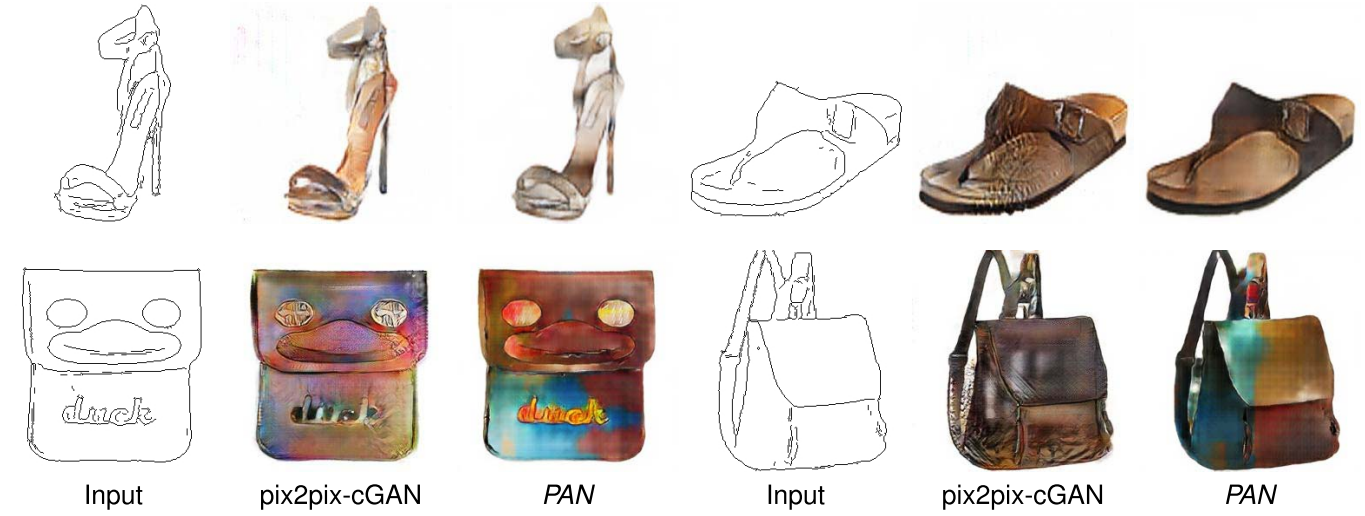

Fig. 9. Comparison of transforming the object edges to corresponding images using the pix2pix-cGAN with the proposed PAN. Given the edges (leftmost), the generated images of shoes and handbags are listed on the rightside.

a serial of image-to-image transformation tasks, such as translating the object edges to its photos, semantic labels to scene images, gray images to color images, *etc*. The proposed PAN can also solve the image-to-image transformation tasks performed by pix2pix-cGAN. Here, we implemented some of them and compared with pix2pix-cGAN.

Firstly, we attempted to translate the semantic labels to cityscapes images. Unlike the image segmentation problems, this inverse translation is an ill-posed problem and image transformation network has to learn prior knowledge from the training data. As shown in Fig. 8, given semantic labels as input images, we listed the transformed cityscapes images of pix2pix-cGAN, PAN and the corresponding ground-truth on the rightside. From the comparison, we found the proposed PAN captured more details with less deformation, which led the synthetic images are looked more realistic. Moreover, the quantitative comparison in Table V also indicated that the PAN can achieve much better performance.

Generating real-world objects from corresponding edges is also one kind of image-to-image transformation task. Based on the dataset provided by Isola *et al.* [15], we trained the PAN to translate edges to object photos, and compared its

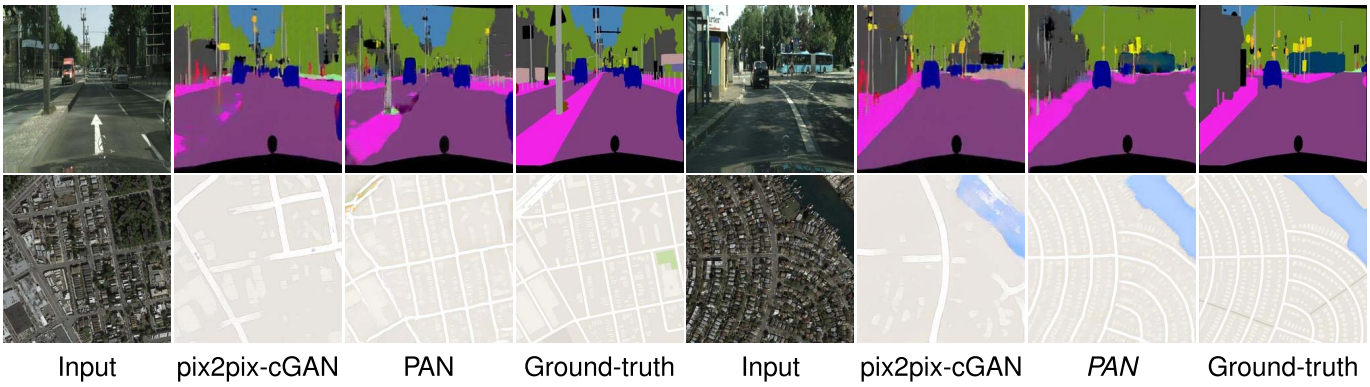

Fig. 10. Comparison of some other tasks using the pix2pix-cGAN with the proposed PAN. In the first row, semantic labels are generated based on the real-world cityscapes images. And, the second row reports the generated maps given the aerial photos as input.

TABLE V
COMPARISON WIHT pix2pix-CGAN

Senmatic labels → Cityscapes imags

	PSNR(dB)	SSIM	UQI	VIF
pix2pix-cGAN	15.74	0.4275	0.07315	0.05208
PAN	**16.06**	**0.4820**	**0.1116**	**0.06581**

Edges → Shoes

	PSNR(dB)	SSIM	UQI	VIF
ID-cGAN	**20.07**	0.7504	0.2724	0.2268
PAN	19.51	**0.7816**	**0.3442**	**0.2393**

Edges → Handbags

	PSNR(dB)	SSIM	UQI	VIF
ID-cGAN	**16.50**	0.6307	0.3978	0.1723
PAN	15.90	**0.6570**	**0.4042**	**0.1841**

Cityscapes images → Semantic labels

	PSNR(dB)	SSIM	UQI	VIF
ID-cGAN	19.46	0.7270	0.1555	0.1180
PAN	**20.67**	**0.7725**	**0.1732**	**0.1638**

Aerial photos → Maps

	PSNR(dB)	SSIM	UQI	VIF
ID-cGAN	26.10	0.6465	0.09125	0.02913
PAN	**28.32**	**0.7520**	**0.3372**	**0.1617**

performance with that of pix2pix-cGAN. Given edges as input, Fig. 9 presented shoes and handbags synthesized by pix2pix-cGAN and PAN. At the same time, the quantitative results over the test set were shown in the Table V. Observing the generated object photos, we think that both pix2pix-cGAN and PAN achieved promising performance, yet it's hard to tell which one is better. Quantitative results are also very close, PAN performed slightly inferior to pix2pix-cGAN on the PSNR measurement, yet superior on other quantitative measurements.

In addition, we compared PAN with pix2pix-cGAN on tasks of generating semantic labels from cityscapes photos, and generating maps from the aerial photos. Some example images generated using PAN and pix2pix-cGAN and their corresponding quantitative results were shown in Fig. 10 and Table V, respectively. To perform these two tasks, the image-to-image transformation models are asked to capture the semantic information from the input image, and synthesize the corresponding transformed images. Since pix2pix-cGAN employed the pixel-wise and generative adversarial losses to training their model, it may hard to capture perceptual information from the input image, which causes that their results are poor, especially on transforming the aerial photos to maps. However, we can observe that the proposed PAN can still achieve promising performance on these tasks. These experiments showed that the proposed PAN can also effectively extract the perceptual information from the input image.

Overall, in the cityscapes dataset (Fig. 8), semantic labels correspond to different kinds of objects, such as vehicle, road, tree, *etc*. Meanwhile, objects in the same category usually share some common patterns and features. Similarly, in the aerial2map dataset (Fig. 10), both aerial images and maps have their own shared patterns, which is beneficial for learning the mapping relation between them. In contrast, in the edges2images task (Fig. 9), given a handbag sketch, it can correspond to hundreds of kinds of outputs with different colors, textures, *etc*. Therefore, compared with Fig. 8 (cityscapes) and Fig. 10 (aerial2map), transformation relations in Fig. 9 (edges2images) are more difficult and challenging, which leads to relatively small visual improvement.

E. Extension to Unpaired Image Translations

As discussed in section II, besides learning paired image-to-image transformations, some works [16]–[18] investigated cross-domain image translations and performed image translations in absence of paired examples. The proposed perceptual adversarial loss aims to continually explore and minimize the discrepancy between perceptual features of generated images and that of the target images. However, in unpaired image translation tasks, the target image corresponding to a generated image is unknown. Alternatively, given training samples $\{x_i \in \mathcal{X}\}_{i=1}^{N}$ and $\{y_i \in \mathcal{Y}\}_{i=1}^{N}$ in two domains, we calculate the discrepancy of mean features on two domains,

$$\ell_{percep}^{D,j} = ||\mathbb{E}_{\mathcal{Y}} d_j(y_i) - \mathbb{E}_{\mathcal{X}} d_j(G(x_i))||, \tag{7}$$

where $G(\cdot)$ and $d_j(\cdot)$ represent the output of generators and the representation on the jth hidden layer of discriminators in the CycleGAN framework, respectively.

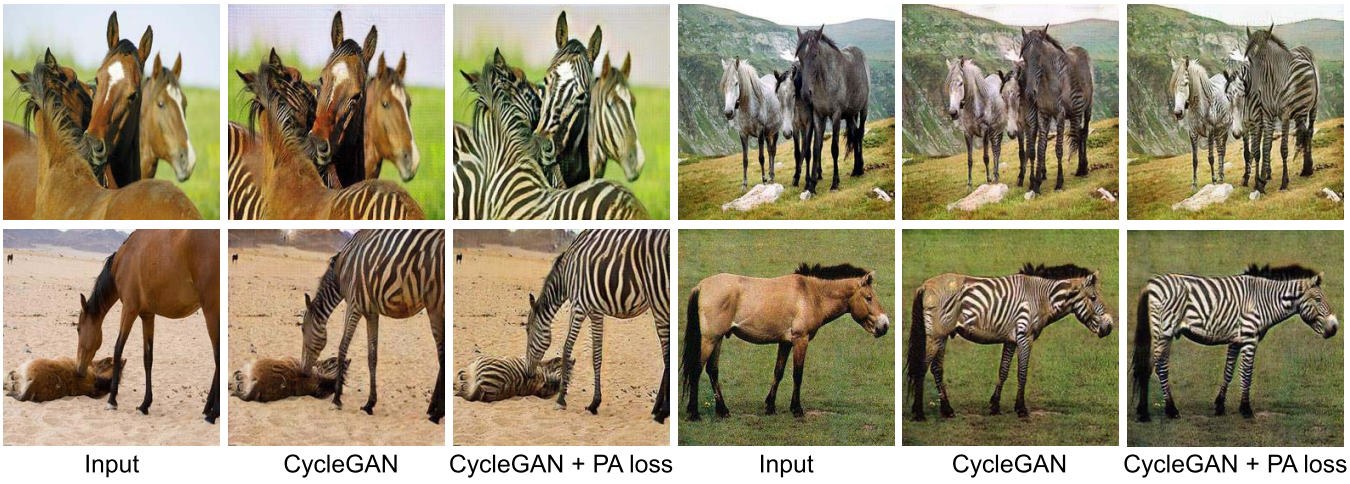

Fig. 11. Introducing the proposed perceptual adversarial loss to CycleGAN framework and attempting to perform unpaired image translation (horses↔zebras). Given horse images as input, trained models (both CycleGAN and 'CycleGAN+Perceptual adversarial loss') aim to generate corresponding zebras images.

In this section, we perform the unpaired image translation task, horse↔zebra, and qualitatively report some generated results. For a fair comparison, we adopted default settings (and codes) from CycleGAN, and utilized the 3rd and 4th hidden layers of the discriminator to measure the perceptual adversarial loss. In addition, hyper-parameters $\lambda_3 = \lambda_4 = 0.5$, $m = 0.1$, and batch size of 4 were used. As shown in Fig. x, through considering the perceptual similarity in unpaired image translations, the model performance was more or less improved. Although this work mainly focuses on exploring the perceptual features between paired images (the generated image and its ground-truth), we demonstrate the possibility of improving the performance of unpaired image translations through measuring the perceptual similarity between different image domains.

V. Conclusion

In this paper, we proposed the perceptual adversarial networks (PAN) for image-to-image transformation tasks. As a generic framework of learning mapping relationship between paired images, the PAN combines the generative adversarial loss and the proposed perceptual adversarial loss as a novel training loss function. According to this loss function, a discriminative network D is trained to continually and automatically explore the discrepancy between the transformed images and the corresponding ground-truth images. Simultaneously, an image transformation network T is trained to narrow the discrepancy explored by the discriminative network D. Through the adversarial training process, these two networks are updated alternately. Finally, experimental results on several image-to-image transformation tasks demonstrated that the proposed PAN framework is effective and promising for practical image-to-image transformation applications.

Acknowledgment

The authors would like to thank the handling associate editor Dr. Catarina Brites and all anonymous reviewers for their positive support and constructive comments for improving the quality of this paper.

References

[1] M. Elad and M. Aharon, "Image denoising via sparse and redundant represent. Over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.
[2] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. 27th Annu. Conf. Comput. Graph. Interactive Techn.* Hoboken, NJ, USA: Wiley, 2000, pp. 417–424.
[3] K. Nasrollahi and T. B. Moeslund, "Super-resolution: A comprehensive survey," *Mach. Vis. Appl.*, vol. 25, no. 6, pp. 1423–1468, 2014.
[4] Q. Luan, F. Wen, D. Cohen-Or, L. Liang, Y.-Q. Xu, and H.-Y. Shum, "Natural image colorization," in *Proc. 18th Eurograph. Conf. Rendering Techn.* Aire-la-Ville, Switzerland: Eurograph. Assoc., 2007, pp. 309–320.
[5] M. W. Khan, "A survey: Image segmentation techniques," *Int. J. Future Comput. Commun.*, vol. 3, no. 2, p. 89, 2014.
[6] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley, "Clearing the skies: A deep network architecture for single-image rain removal," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2944–2956, Jun. 2017.
[7] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.
[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.
[9] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 649–666.
[10] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 415–423.
[11] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2016.
[12] Y. Du, T. Liu, Y. Li, R. Duan, and D. Tao. (2018). "Quantum divide-and-conquer anchoring for separable non-negative matrix factorization." [Online]. Available: https://arxiv.org/abs/1802.07828
[13] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.
[14] T. Miyato and M. Koyama, "cGANs with projection discriminator," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–23.
[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. (2016). "Image-to-image translation with conditional adversarial networks." [Online]. Available: https://arxiv.org/abs/1611.07004
[16] 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. (ICCV)*, Oct. 2017, pp. 2242–2251.
[17] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2868–2876.
[18] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks,"

in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 70, Aug. 2017, pp. 1857–1865.
[19] X. Chen, C. Xu, X. Yang, and D. Tao. (2018). "Attention-GAN for object transfiguration in wild images," [Online]. Available: https://arxiv.org/abs/1803.06798
[20] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. New York, NY, USA: Springer-Verlag, 2016, pp. 694–711.
[21] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 658–666.
[22] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–17.
[23] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[24] H. Zhang, V. Sindagi, and V. M. Patel. (21017). "Image de-raining using a conditional generative adversarial network." [Online]. Available: https://arxiv.org/abs/1701.05957
[25] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4681–4690.
[26] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. (2018). "The unreasonable effectiveness of deep features as a perceptual metric." [Online]. Available: https://arxiv.org/abs/1801.03924
[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning represents by back-propagating errors," *Cognit. Model.*, vol. 5, no. 3, p. 1, 1988.
[28] D. Eigen, D. Krishnan, and R. Fergus, "Restoring an image taken through a window covered with dirt or rain," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 633–640.
[29] T. Ružić and A. Pižurica, "Context-aware patch-based image inpainting using Markov random field modeling," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 444–456, Jan. 2015.
[30] C. Qin, C.-C. Chang, and Y.-P. Chiu, "A novel joint data-hiding and compression scheme based on SMVQ and image inpainting," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 969–978, Mar. 2014.
[31] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.
[32] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.
[33] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 2650–2658.
[34] J. Yang, S. E. Reed, M.-H. Yang, and H. Lee, "Weakly-supervised disentangling with recurrent transformations for 3D view synthesis," in *Proc. 28th Int. Conf. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 1099–1107.
[35] C. Wang, C. Wang, C. Xu, and D. Tao, "Tag disentangled generative adversarial network for object image re-rendering," in *Proc. 26th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017, pp. 2901–2907.
[36] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2366–2374.
[37] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable represent learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 2172–2180.
[38] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–17.
[39] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 214–223.
[40] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5769–5779.
[41] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–26.
[42] C. Wang, C. Xu, X. Yao, and D. Tao. (2018). "Evolutionary generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1803.00657
[43] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–26.
[44] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1–9.
[45] W. Lotter, G. Kreiman, and D. Cox, "Unsupervised learning of visual structure using predictive generative networks," in *Proc. Int. Conf. Learn. Represent. (ICLR) Workshop*, 2016, pp. 1–12.
[46] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. New York, NY, USA: Springer-Verlag, 2016, pp. 597–613.
[47] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Neural photo editing with introspective adversarial networks," in *Proc. 15th Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–15.
[48] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1–13.
[49] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 1–10.
[50] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1–16.
[51] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg, "Transformation-grounded image generation network for novel 3D view synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 702–711.
[52] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–16.
[53] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2011, pp. 2018–2025.
[54] V. Dumoulin and F. Visin. (2016). "A guide to convolution arithmetic for deep learning." [Online]. Available: https://arxiv.org/abs/1603.07285
[55] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 3213–3223.
[56] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 1–8.
[57] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1–9.
[58] J. Bergstra *et al.*, "Theano: A CPU and GPU math compiler in python," in *Proc. 9th Python Sci. Conf.*, 2010, pp. 1–7.
[59] D. Kingma and J. Ba, "ADam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–15.
[60] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[61] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.
[62] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[63] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 184–199.
[64] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–14.

Chaoyue Wang reveived the bachelor's degree from Tianjin University, Tianjin, China, in 2014. He is currently pursuing the Ph.D. degree with the Center of Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney. His research interests mainly include machine learning and computer vision.

Chang Xu received the bachelor's degree in engineering from Tianjin University, China, and the Ph.D. degree from Peking University, China. He is currently a Lecturer of machine learning and computer vision with the School of Information Technologies, The University of Sydney. His research interests lie in machine learning, data mining algorithms and related applications in artificial intelligence and computer vision, including multi-view learning, multi-label learning, visual search, and face recognition. His research outcomes have been widely published in prestigious journals and top-tier conferences. He received fellowships from IBM and Baidu.

Chaohui Wang received the Ph.D. degree in applied mathematics and computer vision from Ecole Centrale Paris, Châtenay-Malabry, France, in 2011. He was a Post-Doctoral Researcher with the Vision Lab, University of California Los Angeles, from 2012 to 2013, and also with the Perceiving Systems Department, Max Planck Institute for Intelligent Systems, Tübingen, Germany, from 2013 to 2014. Since 2014, he has been a Maître de Conférences with Université Paris-Est, Marne-la-Vallée, and a Permanent Researcher with the LIGM Lab, IMAGINE Group, A3SI Team, UMR 8049 CNRS·ENPC·ESIEE·UPEM, Université Paris-Est, France. His research interests include computer vision, computer graphics, image processing, machine learning, and robotics.

Dacheng Tao (F'15) is currently a Professor of computer science and an ARC Laureate Fellow with the Faculty of Engineering and Information Technologies, School of Information Technologies, The University of Sydney, where he is also the Inaugural Director of the UBTECH Sydney Artificial Intelligence Centre. He mainly applies statistics and mathematics to artificial intelligence and data science. His research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and 500+ publications at prestigious journals and prominent conferences, such as IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, ICML, CVPR, ICCV, ECCV, ICDM; and ACM SIGKDD, with several best paper awards, such as the Best Theory/Algorithm Paper Runner Up Award in IEEE ICDM07, the Best Student Paper Award in IEEE ICDM13, the Distinguished Student Paper Award in the 2017 IJCAI, the 2014 ICDM 10-Year Highest-Impact Paper Award, and the 2017 IEEE Signal Processing Society Best Paper Award. He was a recipient of the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award, and the 2015 UTS Vice-Chancellors Medal for Exceptional Research. He is a Fellow of AAAS, OSA, IAPR, and SPIE.