Front. Comput. Sci., 2021, 15(4): 154326
https://doi.org/10.1007/s11704-020-9002-7

RESEARCH ARTICLE

Unpaired image to image transformation via informative coupled generative adversarial networks

Hongwei GE, Yuxuan HAN, Wenjing KANG, Liang SUN (✉)

College of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China

Abstract We consider image transformation problems, and the objective is to translate images from a source domain to a target one. The problem is challenging since it is difficult to preserve the key properties of the source images, and to make the details of target being as distinguishable as possible. To solve this problem, we propose an informative coupled generative adversarial networks (ICoGAN). For each domain, an adversarial generator-and-discriminator network is constructed. Basically, we make an approximately-shared latent space assumption by a mutual information mechanism, which enables the algorithm to learn representations of both domains in unsupervised setting, and to transform the key properties of images from source to target. Moreover, to further enhance the performance, a weight-sharing constraint between two subnetworks, and different level perceptual losses extracted from the intermediate layers of the networks are combined. With quantitative and visual results presented on the tasks of edge to photo transformation, face attribute transfer, and image inpainting, we demonstrate the ICoGAN's effectiveness, as compared with other state-of-the-art algorithms.

Keywords generative adversarial networks, image transformation, mutual information, perceptual loss

1 Introduction

Image-to-image transformation aims to translate available source images into desired targets. The problem has attracted much attention with its wide applications in many tasks including de-noising [1–3], super resolution [4, 5], inpainting [6, 7], colorization [8]. etc. It is difficult since most tasks involve generating target images by utilizing degraded or corrupted source images [9]. Thus ensuring the quality of the target images, as well as keeping the consistency between the source and target images are ongoing challenges.

The progress on image-to-image transformation has been based on convolutional neural networks (CNNs), which follow end-to-end frameworks to map source images into targets by optimizing objectives that evaluate the results in paired data setting [9–11] or unpaired data setting [12–15]. In many applications, the tasks are conducted in unsupervised settings, i.e., paired training examples showing how source images are translated into the corresponding targets are not available. Due to the lack of paired data, the transformation problems become more difficult. From the probabilistic perspective, the challenge is how to infer a joint distribution of images from the marginal distributions in the source and target domain. A generally accepted assumption is that latent paired images in different domains share high-level semantics and exhibit different low-level details. Thus, it is straightforward to enforce weight-sharing constraint on CNNs, with the weight-sharing layers and individual layers encoding the semantics and the details, respectively.

One issue in CNN for image transformation is to design proper loss functions. A straightforward approach is to adopt the pixel-to-pixel loss [16, 17], which minimizes the discrepancy between the generated and the ground-truth images in pixel space. Another approach is to adopt the generative adversarial loss [18, 19], which minimizes the discrepancy between the real data distribution and the generated data distribution. More recently, perceptual loss emerged for minimizing the discrepancy between the generated and ground-truth images at different feature levels [9]. The existing works achieve reasonable results in some tasks. However, coming up with effective losses that yield sharp, realistic images is still an open problem, since the image transformation tasks suffer from blurry results or far from photo-realistic.

Another issue in CNN for image-to-image transformation is how to learn a mapping from source images to targets. The CNN works by learning a mapping from source image domain to a latent space, and then mapping the latent representations to the targets. During the mapping process, it is crucial to transform key properties relating the source images to the target ones. Thus, the problem of how to keep the mutual information among the source images, target images and latent representations consistent needs further studying.

The aforementioned issues imply that there still exists room for researchers to improve their algorithms. In this paper, we propose the informative coupled generative adversarial networks (ICoGAN) in unpaired data setting. The ICoGAN consists of a couple of generators and discriminators, with the parameters in each domain being shared. The algorithm follows the structure of the CoGAN [20], and additionally introduces a

Received January 2, 2019; accepted December 31, 2019

E-mail: liangsun@dlut.edu.cn

mutual information mechanism to keep the consistency among the source images, target images and latent representations. To further improve the quality of generated images at a feature level, we consider the perceptual loss defined by the intermediate layers of the discriminator [9], and extend it to an unpaired data setting by minimizing the difference between the inputs and the generated ones. The contributions of this paper are threefold:

1) We propose the mutual information mechanism in the weight-sharing structure to learn proper common representations of two domains.
2) We combine the mutual information mechanism and the perceptual loss to solve the problems incurred in unsupervised data settings.
3) We evaluate the performance of the proposed ICoGAN on several image transformation tasks and show the importance of each component of the full objective, which validates the effectiveness of the mutual information mechanism and its combination with the perceptual loss.

The rest of the paper is organized as follows: Section 2 summarizes previous work about image transformation and GANs; Section 3 describes the problem formulation; Section 4 describes the proposed model ICoGAN; in Section 5, experiments on MNIST, shoe, handbag, CelebA are conducted to validate the proposed model's effectiveness in image transformation tasks; finally we conclude this work in Section 6.

2 Related work

2.1 Image to image transformation

For image transformation, much progress has been made based on deep convolutional neural networks (CNNs). To solve the image restoration (IR) problem, some works follows an end-to-end setting, e.g., SRCNN [16], MS-LapSRN [21] and DPDNN [22]. SRCNN directly learns a mapping by taking a low-resolution image as the input and produces a high-resolution target. It does not explicitly learn the dictionaries or manifolds for modeling the patch space, but implicitly learn them via hidden layers [16]. To achieve fast and accurate image super-resolution, MS-LapSRN adopts an architecture which progressively reconstructs the residuals of high-resolution images at multiple pyramid levels [21]. DPDNN solves the IR problem by additionally introducing a CNN based denoiser to exploit multi-scale redundancies of images [22]. Another type of CNNs on image transformation introduces a representation learning process. For example, Ma et al. proposed a two stage framework, in which the real embedding features are firstly obtained and then follows the adversarial embedding feature learning [23]. Murez et al. proposed to regularize the extracted representation to perform domain adaptation without training annotations in the target domain [24]. Tran et al. proposed an encoder-decoder structure to learn the representations [25].

The aforementioned CNN based algorithms generally learn a mapping from input images to output images. The problem becomes more difficult when the input images are degraded or when the output images require to exhibit diversities. More recently, the generative adversarial networks (GANs) have emerged as a popular workhorse to solve these problems. By manipulating an adversarial two players' game, GAN learns a distribution of real data. With the powerful adversarial training mechanism, the GAN related algorithms are more effective than the traditional CNNs.

As the tasks require translating input images to targets conditioned on specific properties of the target domain, many works are conducted based on conditional generative adversarial networks (cGANs) [26–30]. For example, Lin et al. proposed to twist two conditional translation models for inputs combination and reconstruction, so that diverse translation results for a fixed input image can be obtained [26]. Li et al. proposed an encoder-decoder architecture, and introduced VGG features and L_1−regularized gradient to obtain a clear image from a hazy one [27]. In [28], Wang et al. proposed a cGAN with multi-scale generator-and-discriminator architecture to synthesize high resolution images. In [29], to solve the cross-view synthesis problem, Regmi and Borji proposed two cGAN architectures, namely X-Fork and X-Seq, both of which learn to produce natural images as well as their semantic segmentation maps. In [30], to solve the image-inpainting problem, examplar information is introduced into cGANs at multiple points. GANs have also been applied to the multi-modal mapping in a unsupervised manner [31–33]. Literature [31] and Literature [32] decompose the image into the content representation and the style representation. By combining the content representation with the random style representation extracted in the style space, they are able to translating the target images into another domain. In [33], an image representation is comprised of both content information which is shared across domains and style information specific to one domain.

It is crucial to pursue two objectives while transforming: (i) preserve the key properties of the source images; (ii) make the details from both domains being as distinguishable as possible. The problem is difficult since the two objectives are interacting with each other. Many existing cGANs operate by embedding a condition term in the generator and the discriminator so that the first objective can be pursued. However, the second objective are often not pursued since they will generate deterministic output given an input. In this paper, we propose an alternative solution, i.e., adopting the InfoGAN structure [34], which learns disentangled representations by maximizing the mutual information between the latent variables and the targets, instead of embedding an arbitrary condition term in the model. Moreover, we adopt a CoGAN structure [20], which uses a couple of generators and discriminators with weight sharing strategy, so that the first objective can be pursued by the weight-sharing-layers, and the second objective can be pursued by non-weight-sharing layers.

2.2 Loss function design of image to image translation

The aforementioned cGANs are developed with distinct applications and the design of loss functions plays an important role for the performance of algorithms. A commonly used approach is to evaluate the target images pixel-wisely, using least squares or L_1, L_2 norm to calculate the difference between the generated and the ground-truth images [10, 16, 17]. The pixel-to-pixel losses are efficient at test time, requiring only a for-

ward pass through the network. More recently, the perceptual losses emerged as a novel measurement for evaluating different semantic levels of images. It encourages the output image to have similar high-level features extracted from different layers of networks. This strategy has found applications in many problems such as feature image super-resolution [35], style transfer [36], texture synthesis [37].

The strategies of pixel-to-pixel loss is effective when paired examples are provided. However, they do not capture perceptual differences between generated and ground-truth images, so they are not free from blurred results or artifacts. Moreover, it is not appropriate for scenarios in absence of paired examples, since we only have independent sets of images from different domains, and there are no paired examples can be used to learn the end-to-end mapping. The strategies of perceptual loss can be applied in an unpaired data setting, but problem of how to utilize the feature layers and extract useful features still remain unanswered [9, 36].

As can been seen from above, the problems of how to minimizing the difference between the generated and the ground-truth images in unpaired data setting, as well as keeping the quality of the generated images requires further studying. The adversarial loss provides an opportunity to generate images in unpaired data setting [12, 13]. And the perceptual loss provides an opportunity to generate more realistic and natural target images. In this paper, we propose to combine the benefits of both perceptual loss and the adversarial loss, i.e., adopting the adversarial loss to solve the unpaired data problem, and combining the perceptual loss to prevent the learned mappings from generating degraded results.

3 Problem formulations

The unpaired image-to-image transformation problem considered in this paper can be described as follows. Denote two image domains X and Y, with training sets $\{x_i\}_{i=1}^{N}$ ($x_i \in X$) and $\{y_j\}_{j=1}^{M}$ ($y_j \in Y$), and denote the image distribution as $x \sim P_x$ and $y \sim P_y$, respectively. Moreover, paired training data (x, y) is not available in the training set. The objective is to learn an underlying mapping f: $X \to Y$, so that for a given input $x \in X$, the corresponding target $y \in Y$ that preserves the key properties of x can be obtained.

Generally, a suitable mapping f is difficult to obtain, since P_x and P_y are marginal distributions in the source and target domain. Let's take the problem presented in Fig. 1 as an example. In Fig. 1, we consider a transformation problem on CelebA dataset that translates black hair face images (domain X) into the blond hair ones (domain Y). The two-dimensional samples in the figure are obtained from the original face image after conducting principal component analysis. The cross points denote the black hair samples in domain X, and the circle points denote the blond hair samples in domain Y. As can be seen from Fig. 1, the examples from domain X and Y are distributed in different regions, which makes the generating of a suitable target y that well preserve the key properties of input x being difficult.

An alternative solution is to adopt an intermediate latent space strategy. As illustrated in Fig. 2(a), we can learn a mapping from domain X to a latent space, and a mapping from the latent space to domain Y, instead of learning a direct mapping

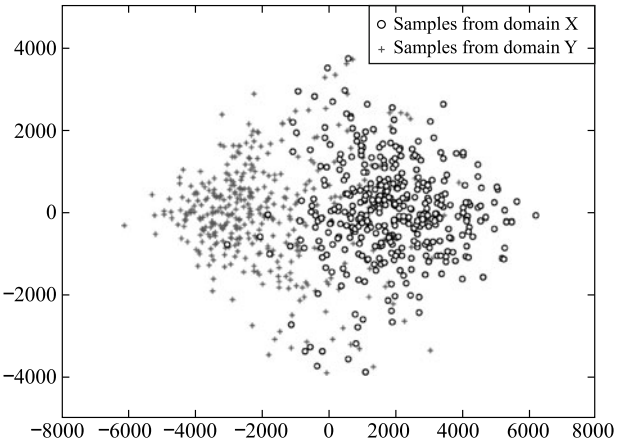

Fig. 1 The distributions data from different domains

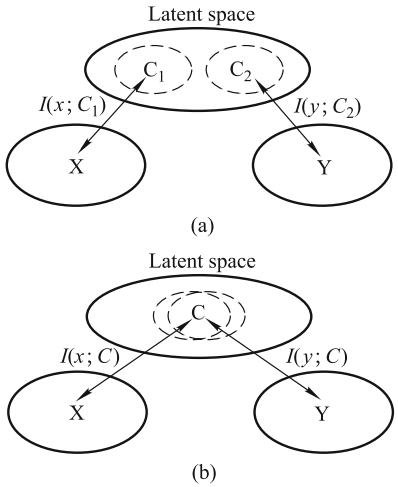

Fig. 2 Semantic diagram of the latent space strategy

from X to Y. In the mapping process, to preserve the key properties of source image x, we can maximize the mutual information $I(x; c_1)$ between x and the mapped latent vector c_1. Similarly, to make the key properties contained in a latent vector c_2 being transferred, we can maximize the mutual information $I(y; c_2)$ between c_2 and the target image y. As illustrated in Fig. 2(b), the problem is then converted to minimizing the discrepancy between c_1 and c_2 to obtain a common latent representation c, which can be easily achieved by deep neural networks.

In this paper, we conduct image transformation based on the above motivations. We construct a generator-discriminator network for source domain. The input image x is fed into the discriminator. By utilizing the mutual information mechanism of infoGAN [34], the discriminator yields a latent vector c_1 that carries the key properties of x and yields a truth/false output. We then construct another generator-discriminator network for target images, with the generator taking another latent vector c_2 as input. The generator yields the target image y. By forcing latent vector c_1 and c_2 converging to a common latent representation c, the indirect mapping from domain X to domain Y can be set up. Moreover, during the above process, the adversarial training mechanism, mutual information acquiring mechanism, and the weight sharing strategy are elaborately designed.

4 Informative coupled generative adversarial networks

In this section, we firstly describe the basic formulations of the generative adversarial networks (GANs) and the coupled generative adversarial networks (CoGANs). And then, we present the details of the proposed informative coupled generative adversarial networks (ICoGAN).

4.1 Models of GANs and CoGANs

Since this paper aims at solving image transformation problems, we describe the formulations of GANs and CoGANs under image generation scenarios. The GAN is designed for learning a distribution of a single domain. It operates by manipulating an adversarial two players' game, i.e., employing a generator and a discriminator, where the generator tries to synthesize instances resembling the real ones, and the discriminator tries to distinguish real instances from synthesized ones. Both the generator and discriminator are realized as multi-layer perceptrons.

Let x be a real image taken from a distribution P_x. Let c be a vector taken from a d-dimensional distribution P_c. Let G and D be the mappings of the generator and discriminator, respectively. The generator takes c as input, and tries to generate an image $G(c)$ that simulates a real one. The discriminator takes x or $G(c)$ as input, and tries to distinguish whether it is a real image or a simulated one. The generator and the discriminator are trained jointly by playing a min-max two-players game. The objective function is defined as follows.

$$\min_G \max_D V_1(G, D) = E_{x\sim P_x}[\log D(x)] + E_{c\sim P_c}[\log(1 - D(G(c)))]. \tag{1}$$

The CoGAN is designed for learning a joint distribution of two domains. It employs a pairs of GANs, i.e., GAN_1 and GAN_2, with each being responsible for synthesizing images in one domain. By using GAN_1 and GAN_2 as two subnetworks of the framework, the CoGAN can learn the joint distribution of data and generate correlated pairs of images in each domain [20].

Let x be a real image taken from distribution P_x, and y be a real image taken from distribution P_y. Let c be a vector taken from a d-dimensional distribution P_c. Let G_1, G_2, D_1, D_2 be the mappings of the two groups of generators and discriminators, respectively. The generators G_1 and G_2 take c as input, and try to generate images $G_1(c)$ and $G_2(c)$ that simulates real images of P_x and P_y, respectively. The discriminator D_1 takes x or $G_1(c)$ as input, and discriminator D_2 takes y or $G_2(c)$ as input. They try to distinguish whether their inputs are real images or simulated ones. Based on the idea that pairs of correlated pairs of images in two domains share the same key properties, the GAN_1 and GAN_2 tie a subset of model parameters, so that correlated pairs of images can be synthesized without correspondence supervision. Similar to GAN, the generators and discriminators are trained jointly by playing min-max two-players game. The objective function is defined as follows.

$$\begin{aligned}\min_G \max_D V_2(G_1, G_2, D_1, D_2) = {} & E_{x\sim P_x}[logD_1(x)] + \\ & E_{c\sim P_c}[log(1 - D_1(G(c)))] + E_{y\sim P_y}[logD_2(y)] + \\ & E_{c\sim P_c}[log(1 - D_2(G(c)))].\end{aligned} \tag{2}$$

The CoGAN generates correlated pairs of images given a random input. However, it doesn't transform an image from one domain to another. In the next subsection, we discuss how we realize image transformation tasks.

4.2 Models of ICoGANs

The framework of the ICoGAN, illustrated in Fig. 3(a), is build on CoGANs [20]. Similar to CoGAN, it consists of 4 subnetworks, including two domain image generators G_1 and G_2, two discriminators D_1 and D_2. For cross domain transformation. we extend GAN_1 and GAN_2 into InfoGAN structure [34] by adding extra networks Q_1 and Q_2, which generate d-dimensional latent vectors c_1 and c_2, respectively. Q_1 and Q_2 make c_1 and c_2 carrying the key properties of the input images by a mutual information mechanism [34]. Then c_1 and c_2 are forced to converge to the original input c, which is used as input to generate simulated images. As illustrated in Fig. 3(b), the above process makes the image transformation data flow of $x \to Q_1 \to c \to G_2 \to y$ or $y \to Q_2 \to c \to G_1 \to x$.

In the above framework, we implement the mechanisms of mutual information, weight sharing, and design the loss function for the network training. The details are as follows.

4.2.1 Mutual information mechanism

The basic InfoGAN operates by dividing the d-dimensional input vector c into a incompressible noise and a latent code, and then minimizing the entropy between input c and the generated image $G(c)$. It is able to learn explicitly representations on challenging tasks [34]. In our proposed model, we utilize the slightly modified mutual information term $I(c; G(c))$ to learn a latent representation. We eliminate the random noise and only use the common representation as the input. The objective function can be expressed as.

$$\min_G \max_D V_3(G, D) = V_1(D, G) - \lambda I(c; G(c)), \tag{3}$$

where G and D are generator and discriminator of the basic GAN, V_1 is the objective function of basic GAN described in Eq. (1), $I(c, G(c))$ is the mutual information between c and $G(c)$, and λ is a hyper parameter that denotes the relative importance of the mutual information $I(c; G(c))$. Generally, it is difficult to obtain the mutual information $I(c; G(c))$ directly, since the posterior probability $P(c|x)$ is difficult to obtain. Thus we estimate it via an auxiliary distribution $Q(c, x)$. The derivation is as follows.

$$\begin{aligned} I(c, G(c)) &= H(c) - H(c|G(c)) \\ &= E_{x\sim G(c)}[E_{c'\sim P(c|x)}[\log P(c'|x)]] + H(c) \\ &= E_{x\sim G(c)}[D_{KL}(P(\cdot|x)||Q(\cdot|x)) \\ &\qquad + E_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \\ &\geqslant E_{x\sim G(c)}[E_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c). \end{aligned} \tag{4}$$

With $Q(c|x)$ approximating $P(c, x)$, we obtain the following variational estimation of mutual information.

$$\begin{aligned} L_I(G, Q) &= E_{x\sim G(c), c\sim P(c|x)}[\log Q(c, x)] + H(c) \\ &\leqslant I(c, x). \end{aligned} \tag{5}$$

Thus optimizing $L_I(G, Q)$ with respect to G and Q is equivalent

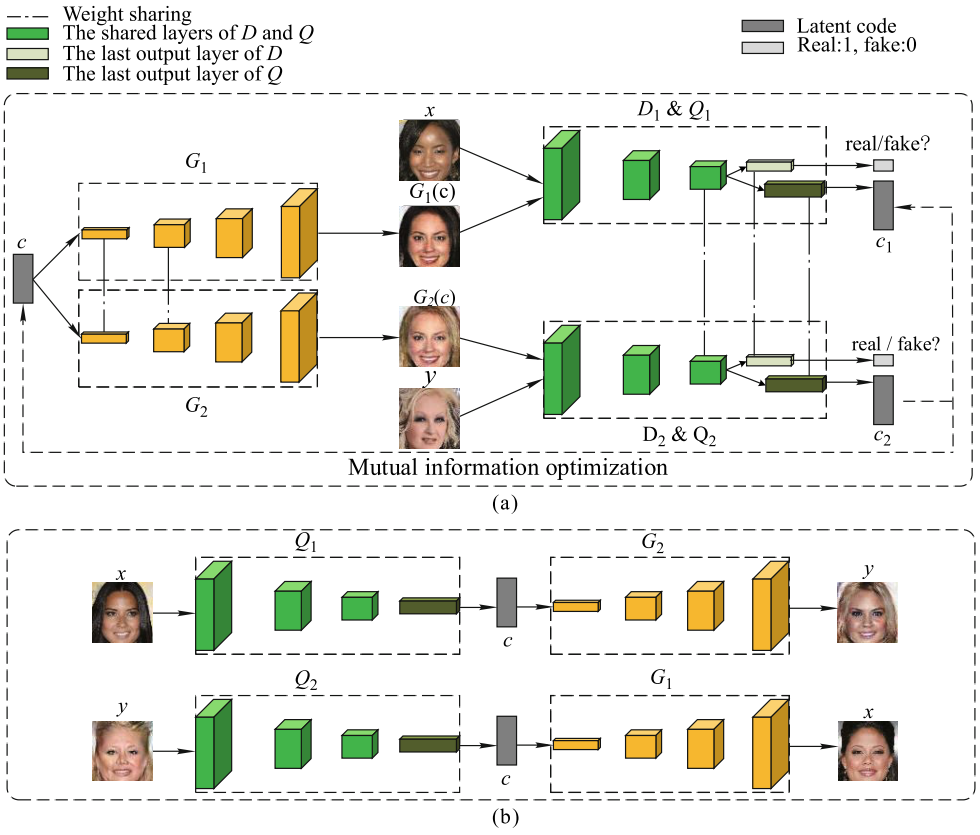

Fig. 3 The framework of the informative coupled generative adversarial networks

to optimizing the mutual information $I(c, G(c))$. The optimization problem in Eq. (3) can be rewritten as.

$$\min_G \max_D V_3(G, D) = V_1(D, G) - \lambda L_I(G, Q). \tag{6}$$

In the implementation, the same strategy is adopted as InfoGAN, which parameterizes a auxiliary distribution Q as a neural network to output the posterior probability $Q(c|x)$ [34]. By optimizing the term $I(c; G(c))$ by network Q, we not only compel the latent code and its generated data to contain the same key properties, but also obtain an inverse mapping from the generated data space to its latent space.

4.2.2 Weight sharing mechanism

With the weight-sharing strategy between generators G_1 and G_2, we ensure that the corresponding generated images share similarity while maintaining their own attributes. Discriminators D_1 and D_2 are responsible for distinguishing images in corresponding domains, which ensures the image realism in each domain. Through the adversarial training and the mutual information maximization, the corelated images in two domains share the same latent code. At the testing stage of the model, any image in one domain can be fed into its discriminator, after which the latent code c can be obtained. With the latent code c fed into the other generator, the model can produces the corresponding image in the target domain.

4.2.3 Design of loss functions

In order to ensure the consistency between the ground truth images denoted by x and corresponding generated ones which are obtained through $G(c)$, we additionally use perceptual loss as a part of the objective. We formulate the pixel-wise discrepancy between the output of the corresponding layers of the paired discriminators as the perceptual loss. Formally, the perceptual loss is defined as follows:

$$P_i = \|D_{1_i}(x) - D_{1_i}(G_1(c_1))\|_2 + \|D_{2_i}(y) - D_{2_i}(G_2(c_2))\|_2, \tag{7}$$

where x and y is the ground truth data, $G_1(c_1)$ and $G_2(c_2)$ are the reconstructed data and $D_{1_i}(\cdot)$, $D_{2_i}(\cdot)$ denote the output of the i_{th} layer of discriminator D_1, D_2 respectively.

Combining the perceptual loss defined in Eq. (7), we formulate the full objective function of our framework as:

$$\begin{aligned}\min_G \max_D V_4(G_1, G_2, D_1, D_2) = V_2(G_1, G_2, D_1, D_2) - \\ \lambda_1 I(c, G_1(c)) - \lambda_2 I(c, G_2(c)),\end{aligned} \tag{8}$$

$$V_{P_1} = \min_G \sum_i \lambda_i P_i, \tag{9}$$

$$V_{P_2} = \min_D [m - \sum_i \lambda_i P_i]^+. \tag{10}$$

The full objective consists of two parts, i.e., V_4 and V_P, which are optimized jointly.

The framework of the proposed ICoGAN is illustrated in Fig. 3. In ICoGAN, a latent code c is fed into both generators G_1 and G_2. x, y are data randomly sampled from real data distribution in domain X and domain Y, respectively. And $G_1(c)$, $G_2(c)$ are data generated by G_1 and G_2, respectively. As shown in the framework, both outputs c_1 and c_2 from the Q_1 and Q_2 are trained to converge to the initially sampled latent code. By such

a process, we avoid the direct mapping between two domains in the data space. Instead, we firstly obtain the latent code c_1 through Q_1, and then use it as a bridge to reach the other domain in the data space by feeding c_1 into G_2.

We note that during the training stage, training data exists or is organized in unpaired data setting. Since we merely randomly draw samples from real data distribution in each domain and feed them to the corresponding discriminator, the training is totally in an unsupervised setting.

5 Experiments

To evaluate our framework, first we validate the effectiveness of the proposed inverse mapping method through experiment on MNIST. Also we conduct experiments on three sets of image transformation tasks, namely edges to photos transformation, face attribute transformation, face inpainting transformation. To further show the model's effectiveness, we compare the results from the proposed model with several state-of-the-art algorithms, i.e., BiGAN [38], CycleGAN [12], UNIT [13] and DiscoGAN [14].

For the network architecture, the network consists of five layers for the discriminators and five layers for the generators. The first three layers of the generator are shared and the last three layers of the discriminator are shared, which enables the framework to learn the joint distribution of the data. The mutual information optimization part is implemented using fully connected layers at the last layer of the discriminator.

We train our framework for 15 epochs with a batch size of 64. We use ADAM as the optimizer for training and set the learning rate and momentum to 0.0002 and 0.5, respectively.

5.1 Evaluation metrics

To present the performance of the proposed ICoGAN's performance on image transformation tasks, we visualize the transformed results. To evaluate our model quantitatively, we adopt the structural similarity index (SSIM) [39], Peak Signal to Noise (PSNR) as evaluation metrics.

SSIM: SSIM measures the similarity between the generated image and the original ground truth. Supposing I_x is the generated image and I_y is the ground truth, the SSIM between I_x and I_y is given by:

$$\mathrm{SSIM}(I_x, I_y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{11}$$

where μ_x, μ_y are the mean pixel value of I_x and I_y respectively, σ_x^2, σ_y^2 are the variance of I_x and I_y respectively, and σ_{xy} is the covariance between I_x and I_y. The bigger the SSIM value is, the closer the generated image is to the ground truth one and the SSIM is non-negative and no larger than 1.

PSNR: PSNR is a criterion measuring the image quality, which is defined as:

$$\mathrm{MSE} = \frac{\sum_{n=1}^{FrameSize} (I_n - P_n)^2}{FrameSize}, \tag{12}$$

$$\mathrm{PSNR} = 10 \times \log \frac{255^2}{\mathrm{MSE}}, \tag{13}$$

where MSE is the mean-square error, I_n is the nth pixel value of the original image, P_n is the nth pixel value after being processed by the model. $FrameSize$ is the product of image's length, width and channel number. The bigger the PSNR, the better the image quality.

5.2 Correctness of the inverse mapping learning method

To validate the effectiveness of the inverse mapping method proposed in Section 2, we conduct image reconstruction task on MNIST to illustrate whether the representation learned by the model denotes the correct semantic meaning. In this experiment, we use a group of generator and discriminator. In such experiment setting, ensured by the perceptual loss and the mutual information, our goal is to learn proper latent representation of the data space, namely the reconstructed digit images should have the same digit type as the input digit images. At the test stage, to perform the reconstruction process, the ground truth image is input to the Q network to obtain the latent code, then the latent code is fed into the generator to decode the latent code into instances in the data space.

Figure 4 illustrates the reconstruction results of the proposed model ICoGAN and the comparison with BiGAN [38]. The first row presents the ground truth image. The second row presents the reconstruction results of the ICoGAN using the representation obtained from the inverse mapping method. The third row presents the reconstruction results of the BiGAN. From the visualization we can see ICoGAN generally outperforms BiGAN since the digit type of several output of the BiGAN are incorrect compared with the ground truth. On the other hand, ICoGAN outputs the images with the correct digit type compared with the ground truth. This indicates that our proposed model is able to learn the latent representation of the data space correctly.

5.3 Edges to photos transformation

We use the shoe images from UT ZAPPOS as the ground truth shoe photo dataset [40] and the handbag images from Amazon as the ground truth handbag photo dataset [41]. After HED edge detection [42], we obtain the corresponding edges. The edge images are taken as the domain X data and the shoe or handbag photos are taken as the domain Y data. We shuffle the data and obtain unpaired data. At the test stage, we feed the edge image to its discriminator to obtain its latent code c, then c will be fed into the generator in the real-photo domain, by which a real shoe photo or a real handbag photo will be produced. The visualized results on edge to shoe or handbag transformation tasks are illustrated in Fig. 5. As shown in Fig. 5, the results pro-

Fig. 4 Visualization results produced by the proposed framework on the digit reconstruction task on MNIST

Fig. 5 Visualization results produced by the proposed framework on edges to photos tasks, i.e., edge to shoe photo transformation and edge to handbag photo transformation. (a) Edges to shoes transformation; (b) edges to handbags transformation

duced by the ICoGAN are generally realistic compared to the state-of-the-art compared models including CycleGAN [12], UNIT [13] and DiscoGAN [14].

We compare the SSIM and PSNR of the ICoGAN with DiscoGAN, CycleGAN and UNIT on the edge to shoe transformation task. SSIM and PSNR are typical evaluation metrics for image quality. The results are presented in Table 1. When the differences between images are great, SSIM and PSNR can well represent the generated images' quality. However, SSIM and PSNR do not completely agree with the perceptual similarity judged by human. Sometimes images with higher SSIM and PSNR are considered as images of worse quality by human. As shown in Fig. 5, the transformation results produced by the ICoGAN stand comparison with the DiscoGAN, CycleGAN and UNIT although the SSIM and PSNR of the ICoGAN are slightly lower. As shown in Fig. 5, CycleGAN tend to generate shoes in gray or in black meanwhile preserving the shoe outlines well. UNIT is good at generating shoes of diverse colors, however it is not as good at preserving the outlines. Therefore, in terms of color diversity, ICoGAN is comparable to DiscoGAN and UNIT, and is better than CycleGAN. In terms of outline preservation, ICoGAN is better than UNIT, worse than CycleGAN and comparable to DiscoGAN. When the difference of SSIM and PSNR between the compared models is small, the evaluation results needs to be further studied.

We further conduct ablation study on the components of the full objective, whose results are listed in Table 2 and Fig. 6. In Fig. 6 and Table 2, w/o denotes without perceptual loss, Pe loss denotes perceptual loss and M term denotes the mutual information term. From Table 2, we can see that without perceptual loss or mutual information term, the quality of the transformation results has declined. When the perceptual loss is replaced with the conventional pixel loss and the LPIPS loss [43] , the results are also not good, which indicates the role played by perceptual loss is important.

5.4 Face attribute transformation

We use the CelebA dataset to train our framework on face attribute transformation task. CelebA is a large scale face attribute

Table 1 Comparison results on edges to shoes transformation

Models	ICoGAN	CycleGAN	DiscoGAN	UNIT
SSIM	0.6984	0.7454	0.7296	0.7023
PSNR	12.6023	14.3772	12.8115	15.9291

Table 2 Ablation study on edges to shoes transformation

Metrics	w/o Pe loss	w/o M term	Pixel loss
SSIM	0.5018	0.5676	0.3849
PSNR	9.8888	10.6055	9.8974

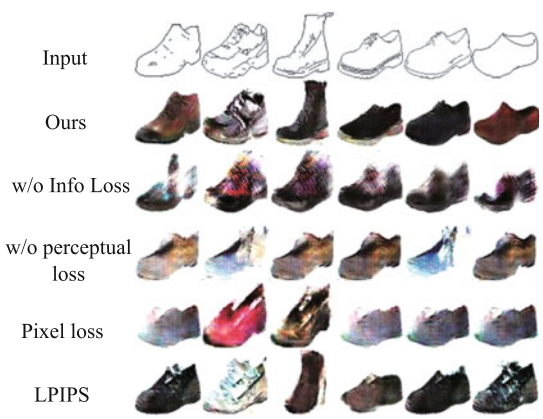

Fig. 6 Visualization of the ablation study

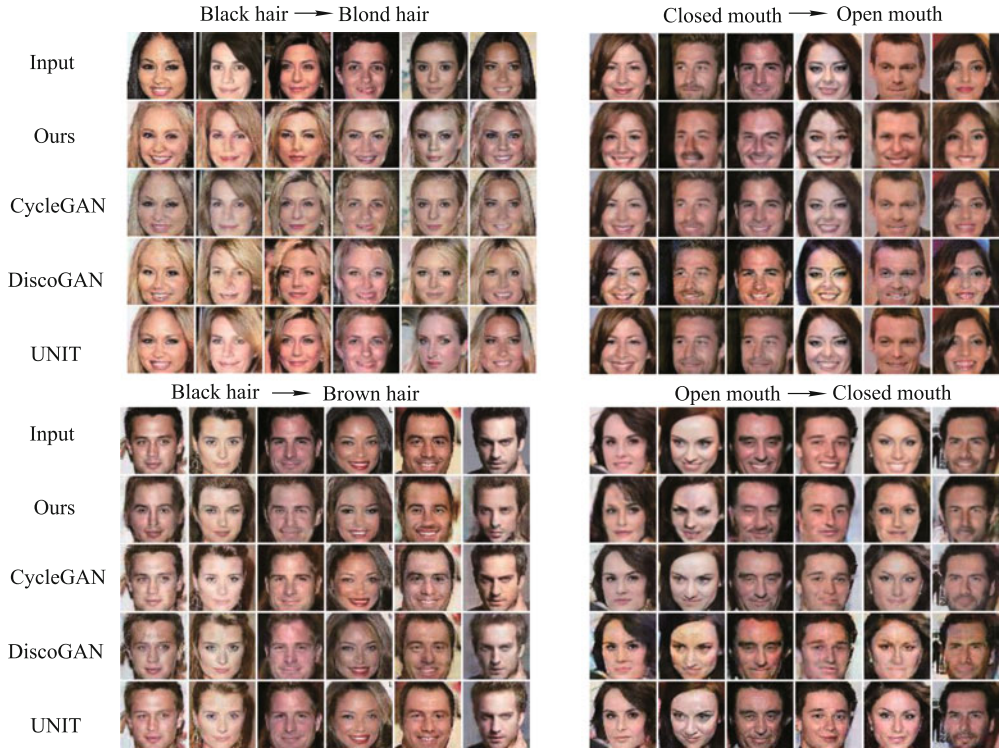

Fig. 7 Visualization results produced by the proposed framework on face attribute transformation task on CelebA including black hair to blond hair transformation, black hair to brown hair transformation, open mouth to closed mouth transformation and closed to open mouth transformation

dataset with over 200K celebrity images, with each having a 40-dimension binary attribute annotation [44]. The annotation consists of attributes such as blond hair, male gender, mouth slightly open, etc. With unpaired training data and in unsupervised setting, we use the ICoGAN to transform images in terms of hair color and mouth condition. All the transformation results on CelebA are visualized in Fig. 7, including black hair to brown hair, black hair to blond hair, open mouth to closed mouth, closed mouth to open mouth.

From the visualization results, we can see that although in unsupervised setting and without paired training data, the framework has successfully transformed images in terms of hair color, mouth condition while retaining the images' attribute. We find that the image generation results are realistic on all four tasks. To evaluate the model, we compare the experiment results with DiscoGAN, CycleGAN and UNIT. Since under such experiment setting, there does not exist a ground truth image in the target domain, the evaluation metrics are not suitable for this task. So we only illustrate the visualization results, which are in Fig. 7. From the visualization comparisons, we find that the results produced by our proposed ICoGAN are comparable to those of DiscoGAN, CycleGAN and UNIT.

5.5 Image inpainting transformation on face

To extend the generality of our model, we also use the CelebA dataset to train our model on image inpainting tasks in an unsupervised setting, where images sampled from real data distribution are considered as domain Y data and images with a quarter of themselves missing as domain X data. The images sampled from the dataset during training are also shuffled to obtain an unpaired experiment setting. By training a pair of weight sharing generative adversarial networks, we can perform image inpainting with the proposed framework.

For the image inpainting transformation tasks, during the test stage, the ground truth images in the target domain exists, so we both list the visualization results in Fig. 8 and quantitative evaluation results in Table 3.

In Table 3, x stands for the ground truth image sampled from the real data while x_p is the inpainted image after the real image with its middle quarter missing is transformed into the full image domain data through the proposed framework. As shown in Table 3, the difference of evaluation metrics between the compared models is small and the inpainted results of the four models are realistic and stable in the attribute retention. As shown in Fig. 8, the ICoGAN preserves the face attribute in the input images and compared to the other three models, the UNIT's results are the best and the ICoGAN's results rank second.

6 Conclusion

In this paper, we present the informative coupled generative adversarial networks (ICoGAN) on image transformation tasks, which operates without paired training data and can relate data from different domains. By appending perceptual loss to the full objective, we encourage the generated images to have similar high-level features with the ground truth ones. Moreover, using mutual information term as a part of the full optimization objective, we assume an approximately-share latent space, which makes the input vector interpretable and the model training efficient. With result visualization and quantitative evaluation

Table 3 Comparison results of PSNR and SSIM on image inpainting

Models	ICoGAN	CycleGAN	DiscoGAN	UNIT
SSIM	0.8486	0.7231	0.7747	0.7474
PSNR	17.8727	20.309	19.4467	21.9634

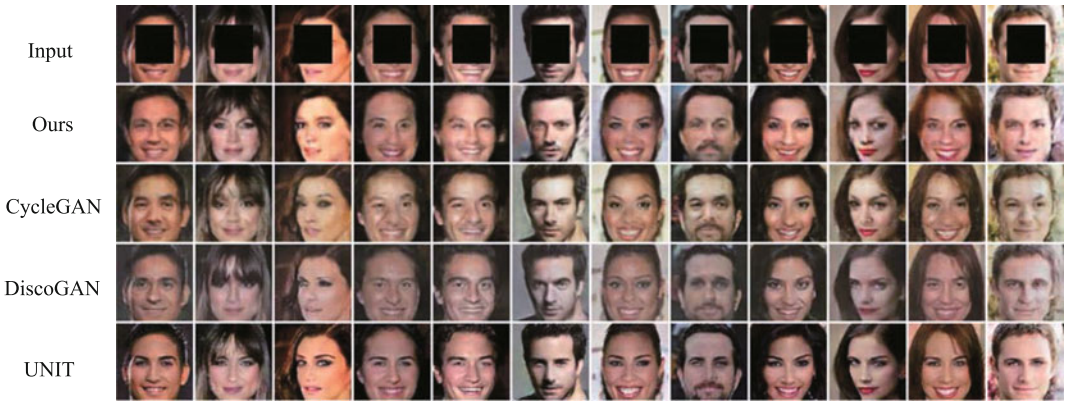

Fig. 8 Visualization results produced by the proposed framework on face inpainting task on CelebA. The first row presents the input data, and the rows from the 2nd to the 5th present the output of ICoGAN, CycleGAN, DiscoGAN, and UNIT, respectively

on both unpaired unsupervised image transformation tasks and supervised image transformation tasks, we demonstrate the proposed model's effectiveness and promising potential for image transformation tasks.

Acknowledgements The authors are grateful to the support of National Key R&D Program of China (2018YFB1600600), the Natural Science Foundation of Liaoning Province (2019MS045), the Open Fund of Key Laboratory of Electronic Equipment Structure Design (Ministry of Education) in Xidian University (EESD1901), the Fundamental Research Funds for the Central Universities (DUT19JC44), and the Project of the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education in Jilin University (93K172019K10).

References

1. Buades A, Coll B, Morel J M. A non-local algorithm for image denoising. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2005, 60–65
2. Elad M, Aharon M. Image denoising via sparse and redundant representations over learned dictionaries. IEEE Transactions on Image Processing, 2006, 15(12): 3736–3745
3. Pan J, Ren W, Hu Z, Yang M H. Learning to deblur images with exemplars. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(6): 1412–1425
4. Cruz C, Mehta R, Katkovnik V, Egiazarian K O. Single image super-resolution based on wiener filter in similarity domain. IEEE Transactions on Image Processing, 2018, 27(3): 1376–1389
5. Huang Y, Li J, Gao X, He L, Lu W. Single image superresolution via multiple mixture prior models. IEEE Transactions on Image Processing, 2018, 27(12): 5904–5917
6. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros A A. Context encoders: feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, 2536–2544
7. Ding D, Ram S, Rodriguez J. Perceptually aware image inpainting. Pattern Recognition, 2018, 83: 174–184
8. Zhang R, Isola P, Efros A A. Colorful image colorization. In: Proceedings of the European Conference on Computer Vision. 2016, 649–666
9. Wang C, Xu C, Wang C, Tao D. Perceptual adversarial networks for image-to-image transformation. IEEE Transactions on Image Processing, 2018, 27(8): 4066–4079
10. Isola P, Zhu J Y, Zhou T, Efros A A. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, 1125–1134
11. Sangkloy P, Lu J, Fang C, Yu F, Hays J. Scribbler: controlling deep image synthesis with sketch and color. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, 5400–5409
12. Zhu, J Y, Park T, Isola P, Efros A A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. 2017, 2223–2232
13. Liu M Y, Breuel T, Kautz J. Unsupervised image-to-image translation networks. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 700–708
14. Kim T, Cha M, Kim H, Lee J K, Kim J. Learning to discover cross-domain relations with generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning. 2017, 1857–1865
15. Huang X, Liu M Y, Belongie S, Kautz J. Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision. 2018, 172–189
16. Dong C, Loy C C, He K, Tang X. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(2): 295–307
17. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017, 39(4): 640–651
18. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. 2015, arXiv preprint arXiv: 1511.06434
19. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Bengio Y. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014, 2672–2680
20. Liu M Y, Tuzel O. Coupled generative adversarial networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016, 469–477
21. Lai W S, Huang J B, Ahuja N, Yang M H. Fast and accurate image super-resolution with deep laplacian pyramid networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(11): 2599–2613
22. Dong W, Wang P, Yin W, Shi G. Denoising prior driven deep neural network for image restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(10): 2305–2318
23. Ma L, Sun Q, Georgoulis S, Gool L V, Schiele B, Fritz M. Disentangled person image generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, 99–108
24. Murez Z, Kolouri S, Kriegman D, Ramamoorthi R, Kim K. Image to image translation for domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, 4500–4509
25. Tran L, Yin X, Liu X. Representation learning by rotating your faces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(12): 3007–3021
26. Lin J, Xia Y, Qin T, Chen Z, Liu T Y. Conditional image-to-image trans-

lation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, 5524–5532
27. Li R, Pan J, Li Z, Tang J. Single image dehazing via conditional generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, 8202–8211
28. Wang T C, Liu M Y, Zhu J Y, Tao A, Kautz J, Catanzaro B. High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, 8798–8807
29. Regmi K, Borji A. Cross-view image synthesis using conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, 3501–3510
30. Dolhansky B, Ferrer C C. Eye in-painting with exemplar generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, 7902–7911
31. Huang X, Liu M Y, Belongie S, Kautz J. Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision. 2018, 172–189
32. Lee H Y, Tseng H Y, Huang J B, Singh M, Yang M H. Diverse image-to-image translation via disentangled representations. In: Proceedings of the European Conference on Computer Vision. 2018, 35–51
33. Ma L, Jia X, Georgoulis S, Tuytelaars T, Van Gool L. Exemplar guided unsupervised image-to-image translation. 2018, arXiv preprint arXiv:1805.11145
34. Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P. Infogan: interpretable representation learning by information maximizing generative adversarial nets. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016, 2172–2180
35. Bruna J, Sprechmann P, LeCun Y. Super-resolution with deep convolutional sufficient statistics. 2015, arXiv preprint arXiv:1511.05666
36. Johnson J, Alahi A, Li F F. Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European Conference on Computer Vision. 2016, 694–711
37. Gatys L, Ecker A S, Bethge M. Texture synthesis using convolutional neural networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. 2015, 262–270
38. Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning. 2016, arXiv preprint arXiv:1605.09782
39. Wang Z, Bovik A C, Sheikh H R, Simoncelli E P. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600–612
40. Yu A, Grauman K. Fine-grained visual comparisons with local learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014, 192–199
41. Zhu J Y, Krähenbühl P, Shechtman E, Efros A A. Generative visual manipulation on the natural image manifold. In: Proceedings of the European Conference on Computer Vision. 2016, 597–613
42. Xie S, Tu Z. Holistically-nested edge detection. In: Proceedings of the IEEE Conference on Computer Vision. 2015, 1395–1403
43. Zhang R, Isola P, Efros A A, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, 586–595
44. Liu Z, Luo P, Wang X, Tang X. Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. 2015, 3730–3738

Hongwei Ge received BS and MS degrees in mathematics from Jilin University, China, and the PhD degree in computer application technology from Jilin University, China in 2006. He is currently a professor and a vice dean in the College of Computer Science and Technology, Dalian University of Technology, China. His research interests are machine learning, computational intelligence, optimization and modeling, computer vision, deep learning. He has published more than 80 papers in these areas. His research was featured in the IEEE Transactions on Cybernetics, IEEE Transactions on Evolutionary Computation, IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, Pattern Recognition, Information Science, etc.

Yuxuan Han received the BS degree from Zhengzhou University, China in 2016, and the MS degree in College of Computer Science and Technology, Dalian University of Technology, China. Her main research interests lie in computational intelligence and machine learning methods.

Wenjing Kang received the BS degree from Northeast University, China in 2016, and the MS degree in College of Computer Science and Technology, Dalian University of Technology, China. Her main research interests are deep learning, machine learning applications such as computer vision and large scale optimization.

Liang Sun received the BE degree in computer science and technology from Xidian University, China, and the MS degree in computer application technology from Jilin University, China in 2003 and 2006, respectively. During 2006–2009, as a DE candidate, he was at College of Computer Science and Technology, Jilin University, China. During 2009–2012, as a DE candidate, he was at Kochi University of Technology (KUT), Japan, as an international student of cooperation between KUT and Jilin University. He received double PhD degree from Kochi University and Jilin University in March, 2012 and June 2012, respectively. He is currently with the College of Computer Science and Technology, Dalian university of technology, Dalian, China. His main research interests lie in machine learning and deep learning.