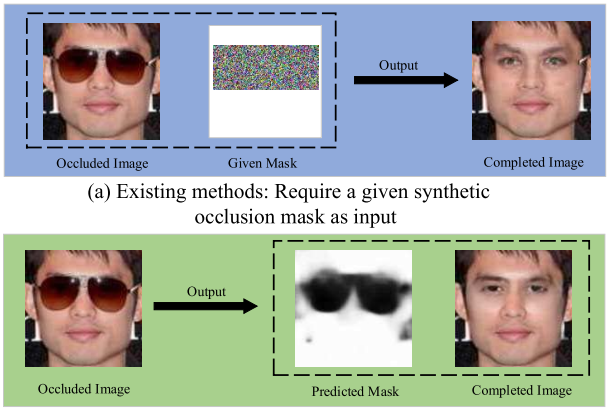
Semi-Supervised Natural Face De-Occlusion

Jiancheng Cai, Hu Han, *Member, IEEE*, Jiyun Cui, Jie Chen, *Member, IEEE*, Li Liu, *Senior Member, IEEE*, and S. Kevin Zhou, *Fellow, IEEE*

***Abstract*— Occlusions are often present in face images in the wild, e.g., under video surveillance and forensic scenarios. Existing face de-occlusion methods are limited as they require the knowledge of an occlusion mask. To overcome this limitation, we propose in this paper a new generative adversarial network (named OA-GAN) for natural face de-occlusion without an occlusion mask, enabled by learning in a semi-supervised fashion using (i) paired images with known masks of artificial occlusions and (ii) natural images without occlusion masks. The generator of our approach first predicts an occlusion mask, which is used for filtering the feature maps of the input image as a semantic cue for de-occlusion. The filtered feature maps are then used for face completion to recover a non-occluded face image. The initial occlusion mask prediction might not be accurate enough, but it gradually converges to the accurate one because of the adversarial loss we use to perceive which regions in a face image need to be recovered. The discriminator of our approach consists of an adversarial loss, distinguishing the recovered face images from natural face images, and an attribute preserving loss, ensuring that the face image after de-occlusion can retain the attributes of the input face image. Experimental evaluations on the widely used CelebA dataset and a dataset with natural occlusions we collected show that the proposed approach can outperform the state of the art methods in natural face de-occlusion.**

***Index Terms*— Natural face de-occlusion, occlusion-aware, generative adversarial networks, alternating training.**



(a) Existing methods: Require a given synthetic occlusion mask as input

(b) Our method: Do NOT require any occlusion mask as input (The task is more challenging)

Fig. 1. (a) Existing methods usually require a given mask, and filling the mask with Gaussian noise before performing de-occlusion. (b) Our method can jointly perform facial occlusion prediction and image de-occlusion.

I. Introduction

OCCLUSIONS often exist in face images of scenarios such as video surveillance and forensics. The problem of face image de-occlusion is an essential and challenging task for face recognition, attribute learning, face parsing, emotion recognition, etc.

The early methods for image de-occlusion or completion are usually exemplar or inpainting based approaches. These approaches can only recover the missing image regions according to the registered or other non-occluded texture information and cannot work well under scenarios with little surrounding texture information left. For face de-occlusion, optimization based methods were proposed in [2], [3] which can only deal with occlusions of limited size. Recently, deep learning based methods [4], [5] [1] were proposed for face de-occlusion, and reported much better results than the traditional approaches. However, these approaches required paired images (i.e., a face image with artificial occlusion and the corresponding non-occluded face image) for training; such paired images may not be available in real scenarios. In addition, these approaches can only deal with artificial occlusions, i.e., by Gaussian block [1] (see Figure. 1 (a)) or an image of object (e.g., glasses, scarf, cup, etc.), but not the natural occlusions in the wild. Moreover, the existing approaches require a given occlusion mask in order to perform de-occlusion.

In this paper, we propose an Occlusion-Aware Generative Adversarial Network (OA-GAN), to perform weekly-supervised natural face de-occlusion using unpaired natural face images, i.e., the ground-truth non-occluded face image of an occluded face image is not available in training. In addition, the proposed de-occlusion approach does NOT require a given mask of the occlusion.

Manuscript received March 4, 2020; revised June 8, 2020 and August 8, 2020; accepted August 23, 2020. Date of publication September 14, 2020; date of current version October 12, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102501; in part by the Natural Science Foundation of China under Grant 61672496 and Grant 61972217; in part by the Youth Innovation Promotion Association CAS under Grant 2018135; and in part by the Natural Science Foundation of Guangdong Province in China under Grant 2019B1515120049 and Grant 2020B1111340056. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Domingo Mery. *(Corresponding author: Hu Han.)*

Jiancheng Cai and Jiyun Cui were with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China (e-mail: jiancheng.cai@vipl.ict.ac.cn; jiyun.cui@vipl.ict.ac.cn).

Hu Han and S. Kevin Zhou are with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: hanhu@ict.ac.cn; zhoushaohua@ict.ac.cn).

Jie Chen is with the School of Electronics and Computer Engineering, Peking University, Beijing 100871, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: chenj@pcl.ac.cn).

Li Liu is with the College of System Engineering, National University of Defense Technology, Changsha 100190, China, and also with the Center for Machine Vision and Signal Analysis, University of Oulu, 90570 Oulu, Finland.

Digital Object Identifier 10.1109/TIFS.2020.3023793

Our OA-GAN can simultaneously predict the occluded region and recover a non-occluded face image (as shown in Figure 1 (b)). As shown in Figure 3, OA-GAN is composed of a generator and a discriminator. The generator consists of an occlusion-aware module and a face completion module. Given a face image with occlusion, the occlusion-aware module of the generator first predicts a mask of the occlusion, which is input to the face completion module of the generator together with the occluded face image for face de-occlusion. The discriminator contains an adversarial loss for discriminating between real face images without occlusions and the recovered face images by de-occlusion, and an attribute preserving loss ensuring the de-occluded face images retaining the same attributes of input face images. We also design an alternating training method in order to obtain better network convergence.

The main contributions of this work are as follows.

(i) We propose a novel semi-supervised approach for natural face de-occlusion without using either paired face images or manually annotated occlusion masks.

(ii) The proposed approach uses a two-stage generator with new encoder-decoder architectures and joint loss functions to perform joint face occlusion detection and de-occlusion.

(iii) The network can be optimized end-to-end by using an alternate training strategy, leading to better network convergence.

(iv) The proposed approach outperforms the state-of-the-art (SOTA) baselines in natural face de-occlusion in both user study and face identification experiments.

II. Related Work

A. Image Completion and Deocclusion

Image completion is to recover the missing content given an image with partial occlusion or corruption. Early image completion methods usually make use of the information of the surrounding pixels around the occluded region to recover the missing part. Ballester *et al.* [6] proposed to perform joint interpolation of the image gray-levels and gradient directions to fill the corrupted regions. Bertalmio *et al.* [7] proposed a variational approach which is based on joint interpolation from the image gradient and the corresponding gray values to the filling-in areas of missing data in a still image. However, these methods may not work well when the missing area in an image is large or has a significant variance in pixel values. Bertalmio *et al.* [8] automatically filled manually selected regions with information surrounding them based on the fact that isophote lines arriving at the regions' boundaries are completed inside. Criminisi *et al.* [9] proposed a patch-based method to search relevant patches from the non-corrupted region of the image and used them to gradually fill the corrupted regions from outside to inside. While such an algorithm provides better results than previous methods, the patch search process can be very slow. In order to solve this issue, Telea [10] proposed a fast patch search algorithm; however, this method still cannot perform image completion in real-time. Then, Barnes *et al.* [11] found approximate nearest-neighbor matches between image patches to speed up completing the missing regions. In general, the patch-based methods rely on the local information and ignore the holistic context information which is also crucial to image completion.

Recently, convolutional neural networks (CNN) based methods were studied for image completion utilizing the whole image's context information. The essence of this kind of method is to predict the missing part by using all the information of the uncorrupted area. Pathak *et al.* [4] proposed the Context Encoders which can understand the content of the entire image and produce a plausible hypothesis for the missing regions. The proposed network used an encoder-decoder architecture with reconstruction loss and adversarial loss. Yu *et al.* [5] proposed a contextural attention deep generative model-based approach to synthesize the missing regions from coarse to fine, which can explicitly utilize surrounding image features as references. However, the method in [5] requires huge computational resources due to its two-stage process for feature encoding. To solve this problem, Sagong *et al.* [12] proposed a parallel extended-decoder path and modified contextual attention module for semantic inpainting. These methods focused on image completion with regular shapes (e.g., rectangle mask), which may different from the case in real applications. In order to overcome this shortcoming, Liu *et al.* [13] used partial convolution and mask-updating jointly to recover arbitrarily shaped area where the convolution is masked and renormalized to be conditioned on only valid pixels. Besides, Zeng *et al.* [14] proposed a pyramid-context encoder to use the information on different scales to improve the image completion results.

Face completion differs from general image completion in that the structures and the shapes of different persons' faces are very similar, but the individual faces' textures are different from each other. Therefore, the face topological structure should be retained during face completion. Zhang *et al.* [3] proposed to perform face completion by moving meshy shelter on the face, which is effective for repairing a small area of corruption. To handle a large area of occlusion, Li *et al.* [1] proposed a face completion GAN, in which a face parsing loss was introduced to maintain the face topological structure, and both global and local discriminators were used to ensure the quality of the completed face image. This approach reported promising results on the CelebA [15] dataset; however, its effectiveness in repairing low-resolution face images with occlusion is not known, while low-resolution and occlusion may simultaneously present in face images in practice.

Cai *et al.* [16] proposed a multi-task learning approach, named FCSR-GAN, to leverage contextual information across different tasks to perform joint face completion and super-resolution. While FCSR-GAN [16] can deal with both face occlusion and low-resolution, it requires paired face images, i.e., low-resolution occluded face images and their mated ground-truth high-resolution face images. In addition, FCSR-GAN requires manual occlusion mask in order to perform face de-occlusion. By contrast, the proposed approach can perform face de-occlusion without requiring manual occlusion mask, and can achieve natural face de-occlusion when there is no paired naturally occluded face image and mated ground-truth face image for training. The face completion module of this work differs from the face completion module

TABLE I
THE ARCHITECTURE OF THE GENERATOR IN OUR OA-GAN

module	type	patch size / stride	padding	output size
Face occlusion-aware module	Conv. + IN + ReLU	7×7 / 1	3	64×128×128
	2× [Conv. + IN + ReLU]	4×4 / 2	1	256×32×32
	3× Residual Block	3×3 / 1	1	256×32×32
	2× [Deconv. + IN + ReLU] (intermediate face feature)	4×4 / 2	1	64×128×128
	Conv. Sigmoid (mask)	7×7 / 1	3	1×128×128
Face completion module	3× [Conv. + IN + ReLU]	4×4 / 2	1	512×16×16
	3× [Deconv. + IN + ReLU]	4×4 / 2	1	64×128×128
	Conv. + Tanh	7×7 / 1	3	3×128×128

(GFC [1] or Pconv [13]) of FCSR-GAN in that: (a) although the face completion modules of both methods share a common encoder-decoder structure, as shown in Table I below, our face completion module consists of different layers; and (b) the face occlusion-aware module in this work is new, which can predict occlusion efficiently and accurately. By contrast, FCSR-GAN does not contain such an occlusion-aware module, and thus cannot be used to predict the face occlusion area.

B. Face Occlusion Detection

Face occlusion detection aims to detect the facial region that is occluded by other objects. Martinez [17] divided the face images into k local regions and designed a probabilistic method to analyze the occlusions in each region. JunOh *et al.* [18] also divided the face image into a finite number of disjoint local patches and then determined whether each patch belongs to the occluded area or not using a PCA model. Min *et al.* [19] focused on scarf and sunglass detection by dividing face image into two equal components and used Gabor features and PCA and SVM models to determine the occluded area.

The face occlusion detection in our OA-GAN differs from the above approaches in two aspects. Firstly, all the above methods divided the face image into several local regions and processed each region separately, which may handle occlusions with regular shape, but may not work well for occlusions with irregular shapes. The occlusion-aware module in our OA-GAN can obtain pixel-level occlusion masks, which can handle face occlusions with arbitrary shapes. Secondly, our method works in a semi-supervised learning and way without using manual occlusion masks, and thus is useful for practical applications.

C. Face Attribute Conversion

Face attribute conversion is to convert the original attributes of the face into other attributes while maintaining the subject identity. The existing methods of face attribute conversion mainly utilize GAN to build face attribute conversion frameworks. Cycle-GAN [20] and Dual-GAN [21] used a weakly supervised method which treats the conversion of two attributes as a conversion between two sets and then used a discriminator for distinguishing face images from the two sets. GANimation [22] and Self-regularization [23] used the attention mechanism so that the network can accurately locate the facial regions to be modified and achieved promising face attribute conversion results. While these methods can edit face attributes in a small area, it is difficult to complete face images with large occlusion.

In this paper, we focus on recovering non-occluded face images from face images with natural occlusions. Similar to face attribute conversion, we use a semi-supervised approach to build our face recovery network. The network structure of the proposed method is similar to GANimation [22] and Self-regularization [23], which can predict the mask for the area of interest and generate the transformed images. While GANimation and Self-regularization use a two-pathway generator, our method uses a two-stage generator. We will discuss the differences of the two types of generators in the next Section. In addition, compared to these face attribute conversion methods, we have designed new joint loss functions for better face de-occlusion.

III. PROPOSED APPROACH

The goal of this work is to learn a face de-occlusion model in a semi-supervised model. Specifically, we aim to tackle a challenging face de-occlusion task that deals with real-world natural occlusions without having the ground-truth non-occluded face images and the occlusion masks.

Let $\{X, Y\}$ denote the training set which contains face images from two domains, i.e., X denoting the natural face images with occlusion x_i, and Y denoting the natural face images without occlusion y_j. However, for any face image x_i in X with natural occlusion, there is NO mated ground-truth face image y_i in Y. In other words, the training face images are unpaired in terms of occlusion and non-occlusion. Our goal is to learn a mapping between X and Y, i.e., $G : X \rightarrow Y$, so that for any x_i, we can obtain $\hat{y}_i = G(x_i)$, where $\hat{y}_i$ belongs to domain Y (a recovered face image without occlusion).

Such an image de-occlusion task is more challenging than conventional face de-occlusion task, in which paired face images are usually available for training. It is also more challenging than face expression conversion, which is actually an holistic image style change. Specifically, the unique challenges of natural face de-occlusion include: (i) While face de-occlusion methods [1] require given occlusion mask to handle artificial occlusions and focus on de-occlusion of artificial occlusion (as shown in Figure 2 (a)), our face de-occlusion method does not require a given occlusion mask, and can deal with natural face occlusions without having the ground-truth non-occluded face images for training; (ii) Although the expression conversion method GANimation [22] does not require paired data for training, the expression conversion task allows changes of the whole face area as long as the subject identity is retained and the face image looks realistic. Face de-occlusion is a more challenging task, because it requires the facial area without occlusions remain unchanged after de-occlusion; however, existing methods like Cycle-GAN aims to perform holistic style transformation without guarantee of local detail preservation after transformation; (iii) While the expression conversion method [22] takes the cycle structure to

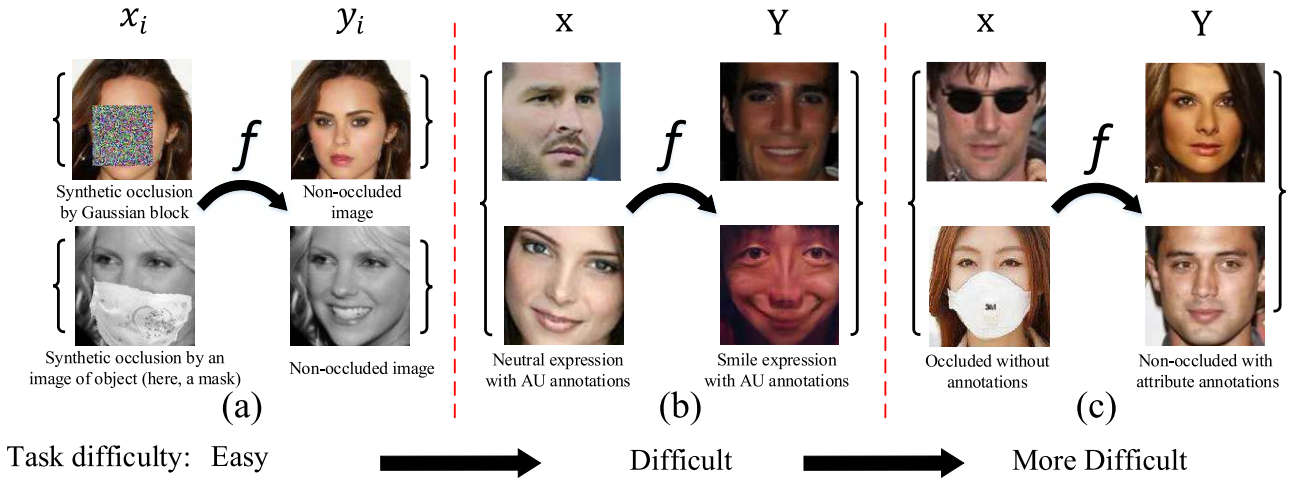


Fig. 2. (a) Paired training data $\{x_i, y_i\}_i$, in which x_i is the face image with artificial occlusion and y_i is its corresponding non-occluded face image. (b) Unpaired training data $\{X, Y\}$ used in the expression conversion task, in which X consists of face images with neutral expression and Y consists of face images with smile expression; there is NO requirement that each natural face image must have a corresponding smile face image of the same subject. (c) Unpaired training data $\{X, Y\}$ in our natural face de-occlusion task, in which X contains face images with natural occlusions and Y contains non-occluded face images but from different subjects than the subjects in X.

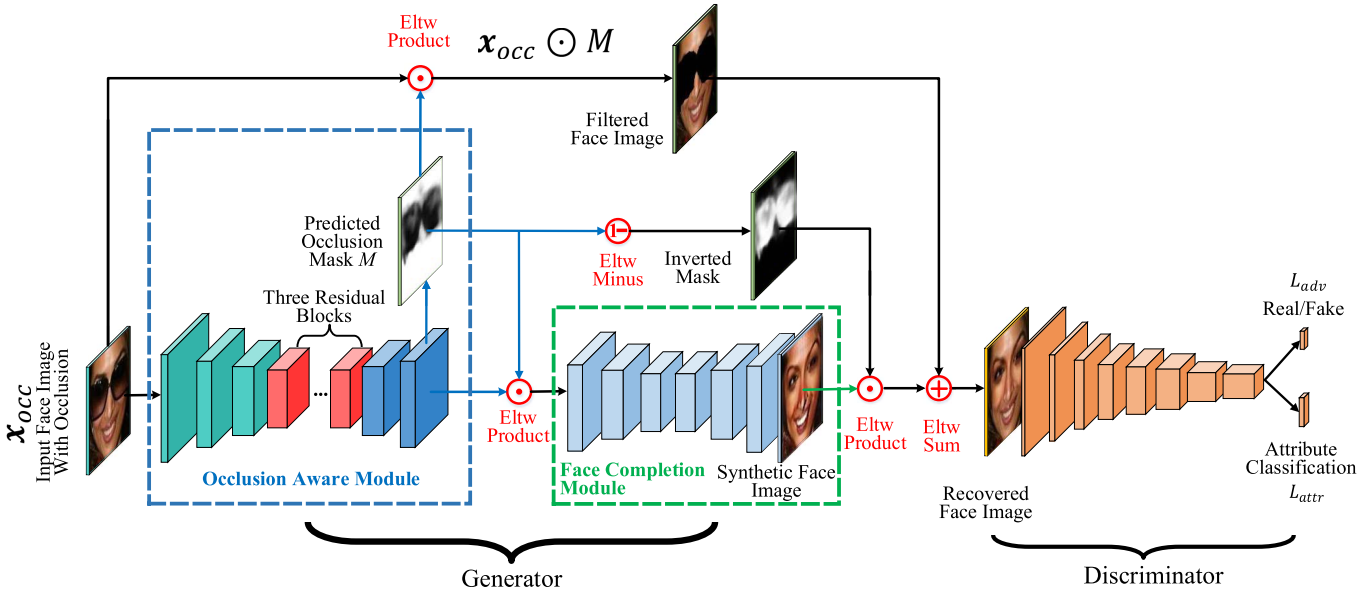


Fig. 3. Overview of our OA-GAN for semi-supervised face de-occlusion without using natural paired images or occlusion masks.

convert the different expression, our face de-occlusion method does not rely on the cycle structure to convert the occluded face images to non-occluded face images; (iv) The amount of annotations used in our method (as shown in Figure 2 (c)) is less than that in [22] (as shown in Figure 2 (b)). For [22], every face image in X has detailed action unit labeling. However, for our method, we only know whether the face images in X is occluded or not. So, natural face de-occlusion task is more difficult than the expression conversion task (as shown in Figure 2).

Besides, our natural face image de-occlusion method also differs from the existing natural face image de-occlusion method [24], which performs face image de-occlusion as an iterative way. A single-pass network inference of [24] usually needs around 1,000 iterations, which requires much more computational cost than our model. In addition, [24] tries to generate a non-occluded face image from a control vector, and the area outside the occlusion mask of the generated face image is expected to be as similar as the same area of the input face image. Different from [24], the second-stage of our generator (i.e., the face completion module) generates a non-occluded face image by using the features of the area outside of the occlusion mask, which is the output of the first-stage of our generator (i.e., the occlusion-aware module).

To this end, we propose an OA-GAN (see Figure 3) to perform face image de-occlusion without having paired natural occluded and non-occluded face images. As shown in Figure 3, our OA-GAN consists of a generator and a

discriminator. The generator jointly performs occlusion prediction and de-occlusion in a cascaded multi-task learning manner [25]. The discriminator follows an auxiliary classifier GAN structure. To obtain better network convergence, we propose a novel training method that alternately feeds paired face images with synthetic occlusions and natural unpaired face images into the generator for semi-supervised learning.

A. Generator

The generator of OA-GAN consists of a face occlusion-aware module and a face completion module, which aims at detecting and restoring from the occlusions, respectively. The face occlusion-aware module has an encoder-decoder architecture consisting of six residual blocks [26], with instance normalization [27] and ReLU layers after every convolution and deconvolution layers in our generator. We use a convolution layer with a sigmoid activation function to regress the occlusion mask. The regressed occlusion mask is a 0-1 filter (0 for occlusion and 1 otherwise) which can be used to keep the texture of non-occluded regions unchanged.

$$Feat_{non\_occ} = M \odot Feat, \tag{1}$$

where $\odot$ represents the element-wise multiplication, M represents the mask. $Feat_{non\_occ}$ represents the occlusion-free feature map which is fed into the face completion module. $Feat$ represents the intermediate face feature of the encoder-decoder network in the occlusion-aware module. The face completion module also follows an encoder-decoder architecture which takes the non-occluded feature map (defined by Equation (1)) as input and generates the texture of occluded regions. Similarly, Instance Normalization and ReLU layers are followed by every convolution and deconvolution layer in face completion module. The detailed generator architecture is shown in Table I.

The output of the face completion module is a synthetic face image including the restored occluded area and the non-occluded area from the input face image. In other words, we only need to restore the occluded area, but keep the non-occluded area unchanged. The final recovered face image is computed as:

$$x_{final} = M \odot x \ + (1 - M) \odot x_{synth}, \tag{2}$$

where x, x_{final} and x_{synth} represent the original occluded face image, the final de-occluded face image and the synthesized face image by our face completion module, respectively.

We should point out that the generator of OA-GAN is essentially different from the generators using in existing methods like GANimation [22] and Self-regularization [23]. GANimation [22] and Self-regularization [23] used a two-path network structure, with one path for mask prediction, and the other path to perform the transformation between two domains. However, the predicted mask is mainly used as a post-processing manner (see Figure 4 (a)). Compared with the two-path generator, the generator of our OA-GAN can leverage the contextual information from the occlusion-aware module (see Figure 4 (b)) to achieve better face completion results.

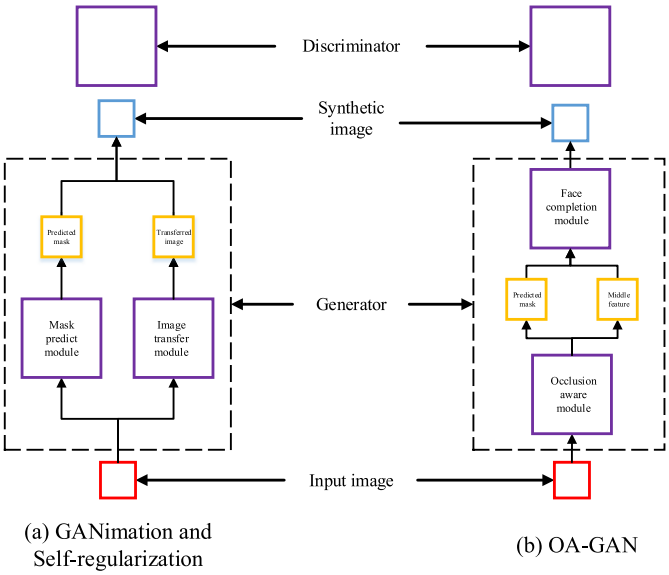


Fig. 4. Different network structures between (a) the generator used in GANimation [22] and Self-regularization [23] and (b) the generator of our OA-GAN.

Specifically, the first stage of the generator (occlusion prediction) will be optimized based on not only the final supervision signal but also the state of the second stage of the generator (face completion). Therefore, the two stages can adapt to each other smoothly. The output of the occlusion-aware module consists of a predicted occlusion mask and an intermediate face feature map. The output by this module is then used as the input of the face completion module, which leads to better non-occluded face image generation results compared to existing methods.

B. Discriminator

The discriminator of OA-GAN plays an auxiliary role in network training. The discriminator is used to determine whether the recovered face is real or fake and whether the recovered face can maintain the attributes contained in the original face image. In our experiments, the supervision signal of attributions can reduce the influence of unbalanced data. For example, in the CelebA dataset, there are significantly less senior people than young people, and less bearded people than people with a beard. We use a total of 10 attributes from CelebA, which are 5_o_clock_shadow, goatee, heavy makeup, male, mustache, no Beard, pale skin, sideburns, wearing lipstick, and young. The structure of our discriminator is similar to Patch-GAN [28], but with modifications of the last layer by adding an attribute classifier. The detailed architecture of the discriminator is shown in Table II. The loss function of the discriminator is defined as follow

$$L_D = \alpha L_{adv} + \beta L_{attr}, \tag{3}$$

where L_{adv} is the adversarial loss [29], and L_{attr} is the mean square error of the attribute between the ground-truth image and the recovered face image. L_{attr} is only used for the paired face images with synthetic occlusions. α and β are two hyper-parameters that balance the influences of the two losses.

TABLE II
THE ARCHITECTURE OF THE DISCRIMINATOR IN OUR OA-GAN

type	patch size / stride	padding	output size
Conv. + LeakyReLU	4×4 / 2	1	64×64×64
5× [Conv. + LeakyReLU]	4×4 / 2	1	2048×2×2
Conv. (adv)	3×3 / 1	1	1×2×2
Conv. (attr)	2×2 / 1	0	10×1×1

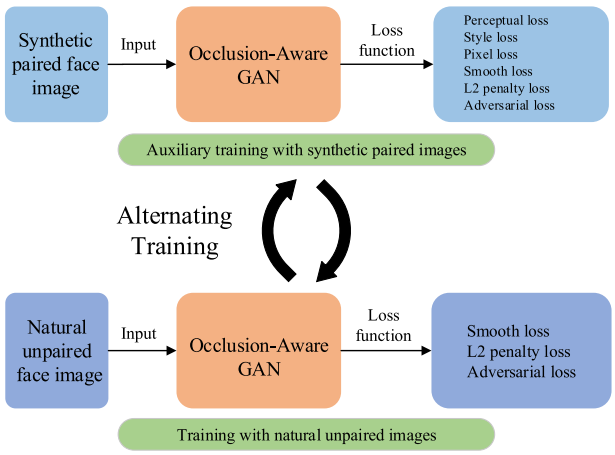


Fig. 5. The diagram of alternating training for our OA-GAN.

C. Alternating Training

The absence of paired face images with and without natural occlusions poses additional challenges to the natural face de-occlusion task. To achieve network convergence in training, we propose an alternating training strategy using different loss combinations in different stages to optimize the whole network. Overall, the alternating training consists of two stages: (i) auxiliary training with paired images with synthetic occlusions, and (ii) training with natural unpaired images (as shown in Figure 5). The former makes a good initialization of the whole network to establish the basic face de-occlusion ability, and the latter expands this ability into the scenarios of unpaired face images with natural occlusions.

Auxiliary Training: In this stage, we establish the basic de-occlusion ability (network initialization) using paired face images with and without synthetic occlusions and a joint loss function containing perceptual loss [30], style loss [13], pixel loss [13], smoothness loss [31], L_2 penalty loss, and adversarial loss [29].

The perceptual loss [30] is used to ensure the low-level pixel values and high-level abstract features as similar as possible between the reconstructed face image and the ground-truth. The perceptual loss is defined as:

$$L_{perceptual} = \sum_{n=0}^{N-1} \|\phi_n(x_{synth}) - \phi_n(x_{gt})\|_1 + \sum_{n=0}^{N-1} \|\phi_n(x_{final}) - \phi_n(x_{gt})\|_1, \quad (4)$$

where the ϕ is the VGG-16 [32] which is pre-trained based on ImageNet, and ϕ_n represents the *n-th* feature maps in VGG model. x_{synth}, x_{final}, and x_{gt} represent the synthetic face image, the non-occluded face image, and the ground-truth image, respectively.

The style loss [13] is used to perform an autocorrelation (Gram matrix) on each feature map and ensure the style unification of the recovered face part and the non-occluded face part. The style loss is defined as:

$$L_{style} = \sum_{n=0}^{N-1} \|K_n(\phi_n(x_{synth})^T \phi_n(x_{synth}) - \phi_n(x_{gt})^T \phi_n(x_{gt}))\|_1 + \sum_{n=0}^{N-1} \|K_n(\phi_n(x_{final})^T \phi_n(x_{final}) - \phi_n(x_{gt})^T \phi_n(x_{gt}))\|_1, \quad (5)$$

where the K_n is the normalization factor $1/(C_n \cdot H_n \cdot W_n)$ for the *n-th* VGG-16 layer, C_n, H_n and W_n are the number, height and width of feature maps, respectively.

The pixel loss [13] is used to ensure the generate face image x_{final} is close to the gound truth x_{gt}, which is defined as:

$$L_{pixel} = \gamma \|(1 - M) \odot (x_{final} - x_{gt})\|_1 + \delta \|M \odot (x_{final} - x_{gt})\|_1, \quad (6)$$

where the γ and δ are scalar factors for balancing different loss functions.

The smoothness loss [31] penalizes the final synthetic face image x_{final} and the mask M if they are not smooth on pixel level, which is defined as:

$$L_{smooth} = \sum_{i,j}^{W,H} (\|x_{final}^{i,j+1} - x_{final}^{i,j}\|_1 + \|x_{final}^{i+1,j} - x_{final}^{i,j}\|_1) + \sum_{i,j}^{W,H} (\|M^{i,j+1} - M^{i,j}\|_1 + \|M^{i+1,j} - M^{i,j}\|_1), \quad (7)$$

where W and H are respectively the width and height of the final synthetic face image x_{final}. The size of mask M is also $W \times H$. $\|\cdot\|_1$ is the L_1 *norm*.

The L2-norm is a penalty term during network training, which can make the predicted occlusion mask as tight as possible; otherwise, some non-occluded area might be predicted as occluded area.

The total loss function for the synthetic paired face images is define as follow:

$$L_{paired} = \lambda_1 L_{perceptual} + \lambda_2 L_{style} + \lambda_3 L_{pixel} + \lambda_4 L_{smooth} + \lambda_5 \|M\|_2^2 + \lambda_6 L_{adv}, \quad (8)$$

where $\lambda_1 - \lambda_6$ are scalar factors for balancing different loss functions and L_{adv} is the adversarial loss.

Training With Natural Unpaired Images: In this stage, we expand the capability of the initial de-occlusion network into the scenario of unpaired face images with natural occlusions by defining a different joint loss function containing smoothness loss, mask L_2 penalty loss, and adversarial loss. The whole loss function is as follow:

$$L_{unpair} = \lambda_4 L_{smooth} + \lambda_5 \|M\|_2^2 + \lambda_6 L_{adv}, \quad (9)$$

Fig. 6. Face occlusion-aware and face completion results by our OA-GAN for face images with artificial occlusion: (a) the input images with artificial occlusion, (b) the masks predicted by OA-GAN, and (c) and (d) the de-occluded face images by OA-GAN and ground-truth face images.

where λ_1, λ_2 and λ_3 are scalar factors balancing different loss functions.

The essence of our alternating network training is to leverage the knowledge from paired face images with synthetic occlusion to assist in the de-occlusoin model learning for natural occlusions. Our alternating training differs from the commonly used two-stage training such as [16] in that: while two-stage training is a step-by-step optimization for different modules of the network using the same data, our alternate training is to use different data to alternately train the entire network. In the training process, we first set the ratio of the synthetic paired face images and the natural unpaired face images to 10:1. Then, we gradually increase the ratio between nature unpaired face images until the ratio of the synthetic paired face images and unpaired natural face images becomes 1:1. This is also helpful in transferring the initial de-occlusion capability into the scenario of unpaired natural face images.

IV. Experimental Results

A. Database

We perform experimental evaluations on the public CelebA dataset [15], which consists of 202,599 face images of 10,177 subjects, with each image containing 40 binary attributes. We treat glasses as one type of natural occlusion, and divide CelebA into two subsets: wearing glasses (13,193 face images) or not (189,406 face images).

We also collect a face dataset with natural occlusions by glasses and respirators, which contains 19,746 face images from CelebA, MAFA [33] and Internet. In our experiments, randomly select 80% of the naturally occluded face images (with either glasses or respirators) and non-occluded face dataset for training, and the remaining 20% face images for testing.

We build the paired face image dataset with synthetic occlusion using more than 640 different types of occlusion objects (eyeglasses, respirators, scarves, etc.). We randomly choose one type of occlusion object and overlay it to a non-occluded face image from CelebA (see some examples in Figure 6). We get more than 100,000,000 pairs of occluded and non-occluded face images in total. Again, we randomly use 80% pairs for training, and the remaining 20% pairs for testing.

B. Training Details

Before training, all face images are aligned based on five facial landmarks (two centers of two eyes, nose tip, and two corners of mouth) provided in CelebA, and are resized to $128 \times 128 \times 3$. For the face images we collected, we locate the five facial landmarks using an open-source algorithm[1]. Although occlusions may affect the accuracy of face landmark detection, the final face de-occlusion results (e.g., the last line of Figure 7) show that the proposed approach still work well.

In addition, in order to make the proposed OA-GAN converge well, we use WGAN-GP [34] to optimize the network. In terms of the optimization, we use the Adam algorithm [35] with an initial learning rate of 10^{-4}. Our loss function in Eq. (8) consists of six items, in which the first four items are designed for face completion and the last two items are designed for occlusion prediction. For the first four items, we have followed [1], [13] to set the hyper-parameters ($\lambda_1 = 0.05$, $\lambda_2 = 120$, $\lambda_3 = 1$, $\lambda_4 = 10^{-3}$). For the last two items, we choose hyper-parameters ($\lambda_5 = -1$, $\lambda_6 = 1$) by considering the scale of the loss values. Although we only use such a simple rule to choose the hyper-parameters, the face de-occlusion results and face recognition results (see Figure 7 and Figure 13) show the effectiveness of the proposed approach.

The baseline methods in our experiments include Cycle-GAN [20], Self-regularization [23], GANimation [22], and [24]. We have used the recommended hyper-parameters in their original papers for all these methods. The divisions

[1] https://github.com/seetaface/SeetaFaceEngine.

Fig. 7. Qualitative comparisons of face de-occlusion results by Cycle-GAN [20], GANimation [22], Self-regularization [23] and our OA-GAN on the CelebA database. (a, i) are the input images to Cycle-GAN, GANimation, Self-regularization and our OA-GAN. (b, d, f, h) and (j, l, n, p) are recovered face images by Cycle-GAN, GANimation, Self-regularization and our OA-GAN, respectively. (c, e, g) and (k, m, o) are predicted occlusion masks by GANimation, Self-regularization and our OA-GAN, respectively.

of training and testing sets for face de-occlusion and face recognition are the same as our approach.

C. Qualitative Comparisons

Qualitative analysis is to compare the visual results, focusing on the reality and rationality of the recovered image. We conducted four different experiments from different aspects. The first experiment is to verify the effectiveness of the proposed OA-GAN in recovering from face images with artificial occlusions (see Figure 6). We can see that the proposed approach can obtain very visually reasonable results compared with the ground-truth face images. The important facial structures and facial characteristics also look very similar to the ground-truth.

The second experiment is to compare the results of our OA-GAN with Cycle-GAN [20], Self-regularization [23], and GANimation [22] in recovering from face images with natural occlusions. From Figure 7, we can see the results of our method are much better than those obtained by Cycle-GAN, Self-regularization, and GANimation. The reason why Cycle-GAN like methods do not work well is that their transformation in Cycle-GAN only involves a change in the whole image style, without guarantee of local detail preservation after transformation. However, in face de-occlusion task, we need to complete the occluded facial area while assuring the non-occluded facial area unchanged after de-occlusion. Besides, although Self-regularization does not rely on a cycle structure, and uses a generator with two-path structure, it does

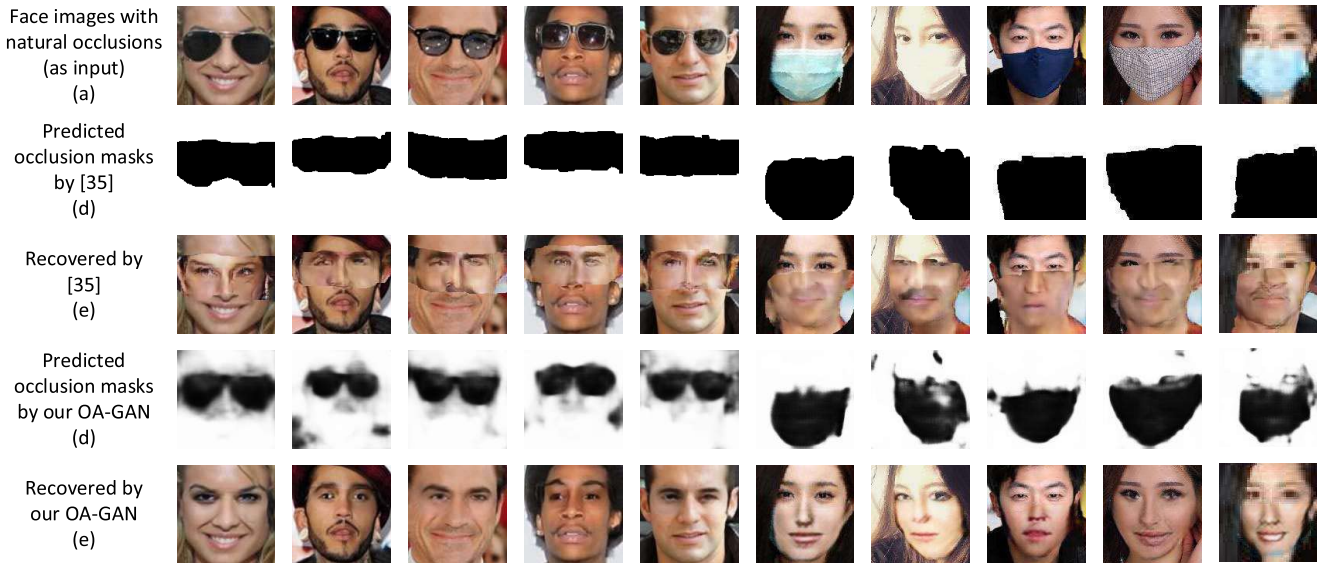


Fig. 8. Qualitative comparisons of face de-occlusion results by [24] and our OA-GAN for several face images with natural occlusion from the CelebA database: (a) the original input face images, (b, d) the predicted occlusion masks by [24] and our OA-GAN, respectively, and (c, e) the recovered face images by and our OA-GAN, respectively.

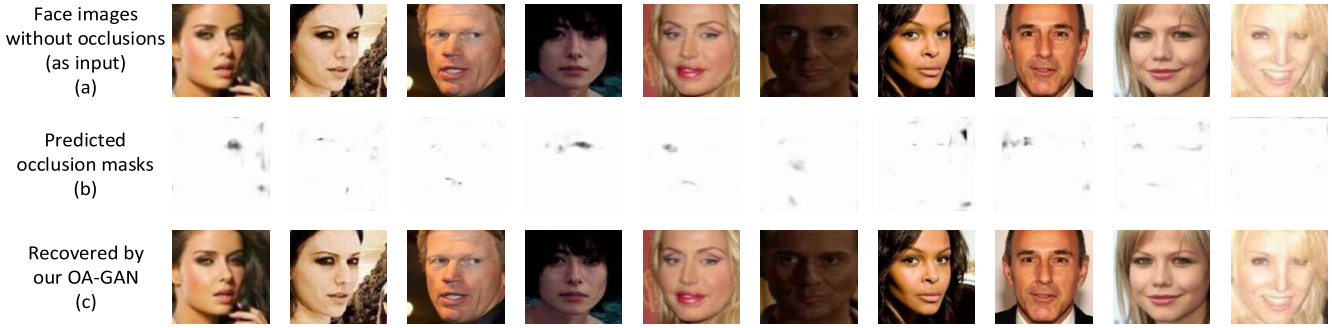


Fig. 9. Results of the proposed OA-GAN in dealing with face images without occlusion: (a) the input face images, (b) the predicted occlusion masks, and (c) the face images after de-occlusion.

not make good use of the contextual information during face completion.

In the third experiment, we compare our OA-GAN with [24], which can perform natural face de-occlusion using DC-GANs. Visual comparisons of the face de-occlusion results by our method and [24][2] are given in Figure 8. We can see our method can generate a non-occluded face image with higher quality than [24]. We consider this is because [24] only tries to generate a non-occluded face image from a control vector to approximate the non-occluded area of the input face image, while our OA-GAN can leverage more contextual information of the whole de-occluded face image in stage-1 to produce a better face de-occlusion result.

Finally, we provide evaluations to see how the proposed OA-GAN can deal with face images without occlusions. As shown in Figure 9, we can see that the proposed OA-GAN predicts very minor occlusion masks for the face images without occlusion. This is a good property because we expect a face de-occlusion algorithm should keep a face image without occlusion unchanged as much as possible.

[2]Since the code of [24] is not publicly available, we reimplemented the method in [24] based on the best of our understanding.

D. Quantitative Comparisons

We perform two experiments to quantitatively evaluate the effectiveness of the proposed approach. In the first experiment, we use two metrics (PSNR and SSIM) to evaluate our method in recovering synthetic face images. PSNR is widely used in image compression area to measure the fidelity of the reconstructed image, and SSIM [36] is a perceptual metric that considers image degradation as a perceived change in structural information, while also incorporating important perceptual phenomena, including both luminance masking and contrast masking terms.

We compare our approach with several baselines including GFC [1], GnIpt [5], Pconv [13] and CSA [37]. For fair comparisons, we use the same synthetic paired face images shown in Figure 10 to conduct the experiment. For our OA-GAN, we let the model predict the position of occlusions by itself. For GFC, GnIpt, Pconv and CSA, since these methods require a occlusion mask as input, we use the mask predicted

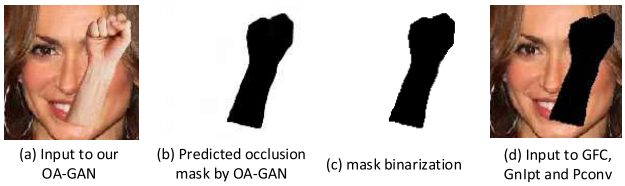


Fig. 10. (a) the input face image with artificial occlusion, (b) predicted occlusion mask by our OA-GAN, (c) the binary mask image for (b), and (d) the input to baseline methods GFC, GnIpt, Pconv and CSA, which is the multiplication result of (a) and (c).

TABLE III
QUANTITATIVE EVALUATIONS OF THE PROPOSED APPROACH AGAINST THE STATE-OF-THE-ART METHOD REQUIRING OCCLUSION MASK AS INPUT (GFC [1], GNIPT [5], PCONV [13], AND CSA [37]) IN TEAMS OF PSNR (dB) AND SSIM

Method	GFC	GnIpt	PConv	CSA	Proposed
PSNR(dB)	19.96	20.13	22.18	22.71	22.61
SSIM	0.718	0.725	0.768	0.794	0.787

by our model as their input. Therefore our method needs to predict the occluded region and then recover the face images, but GFC, GnIpt, Pconv and CSA only need to recover the face images using our mask. The PSNR and SSIM of GFC, GnIpt, Pconv, CSA[3], and our OA-GAN are reported in Table III. We can notice that our approach achieves higher PSNR and SSIM than GFC, GnIpt and Pconv, and comparable results with the CSA for face completion with synthetic occlusions. These results suggest that while the existing methods may not work when manual occlusion masks are not available, the proposed OA-GAN can still obtain very reasonable de-occlusion results.

Since face image is one of the most common images for human visual perception, we believe visual quality of the de-occluded face images is still an important aspect for evaluating face de-occlusion algorithms. Therefore, in the second experiment, we performed a user study by asking three participants to select the best one from the face de-occlusion results by our OA-GAN, and three SOTA de-occlusion methods (Cycle-GAN, GANimation, and Self-regularization). The de-occlusion results of 300 face images with natural occlusion are presented to the participants. Each time, the de-occlusion results by four different methods are displayed on screen in a random order to avoid bias of a fixed order, and each participant is ask to select the best one from the four de-occluded face images. The user study results are given in Figure 11. We can see that our method achieves much better results than the SOTA methods in user study, which indicates that face de-occlusion results by our method have better perceptual quality than the SOTA methods. While a perfect quantitative measurement is still not available for face de-occlusion tasks, we observe that SSIM scores of individual methods are more consistent to our user study results.

[3]Since the code for Pconv and CSA are not publicly available, we re-implemented the two methods based on the best of our understanding.

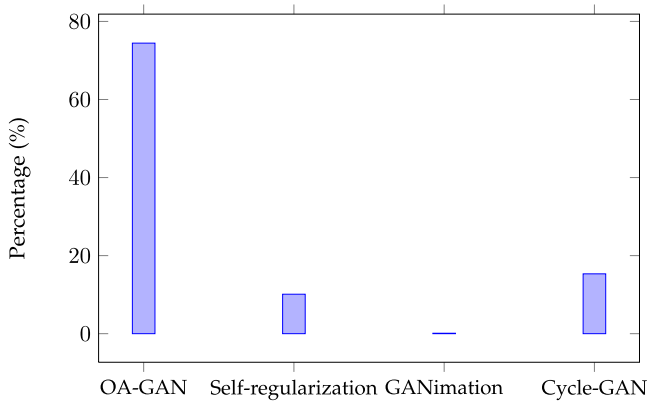


Fig. 11. The percentage of best face de-occlusion results in user study for our OA-GAN, GANimation, Cycle-GAN and self-regularization.

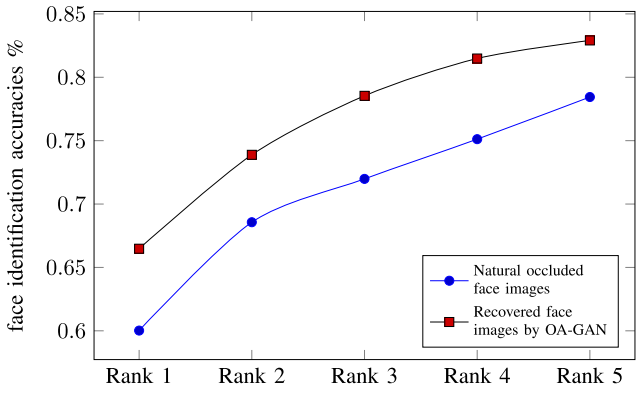


Fig. 12. Rank 1-5 face identification accuracies using natural occluded face images and recovered face images by OA-GAN.

E. Effectiveness for Face Recognition

We also study whether the proposed OA-GAN can improve face recognition when using the recovered face images for face recognition. We have followed the related studies [16], [38], and use the face recognition model trained on de-occluded face images of the training set to recognize the de-occluded testing face images. Such a manner has been found to be effective in mitigating the domain gap between the original images in gallery set and the de-occluded image in probe set. We choose LightCNN-9 [39] as a face recognition model and use two types of face images to train it: (a) original face images in CelebA, and (b) de-occluded face images by OA-GAN. The training, gallery, and probe sets for face identification contain 162,770, 348, and 1,053 face images, respectively[4]. The rank 1-5 face identification rates are shown in Figure 12. We can see that face identification using the recovered face images by our approach can lead to higher accuracy than using the original face images with occlusion. We also visualize some face de-occlusion results by our OA-GAN in Figure 13. Since the ground-truth face images are not available for face images with natural occlusions, we use another face image without occlusion of the same person as a reference of the ground-truth. We can see that the recovered facial areas by our method look very reasonable and realistic, w.r.t. the reference

[4]The sizes of the gallery and probe sets are not very large because it is difficult to find a lot of face images with natural occlusions.

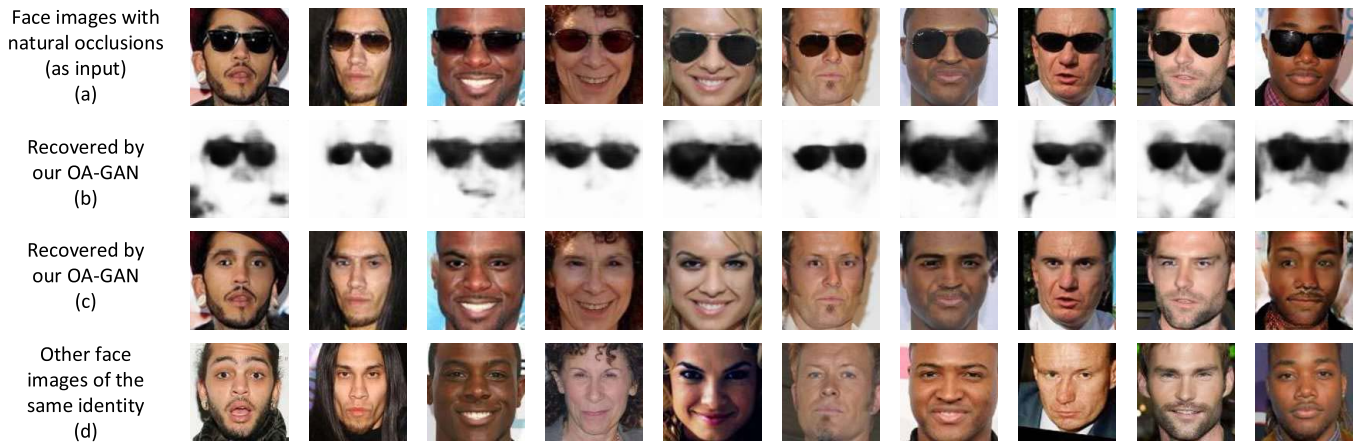


Fig. 13. Comparison between the recovered face images by OA-GAN and another reference ground-truth face image without occlusion of the same person. (a) are the input face images. (b) are the predicted occlusion masks. (c) are the recovered face images by our OA-GAN, and (d) are another face images without natural occlusion of the same person.

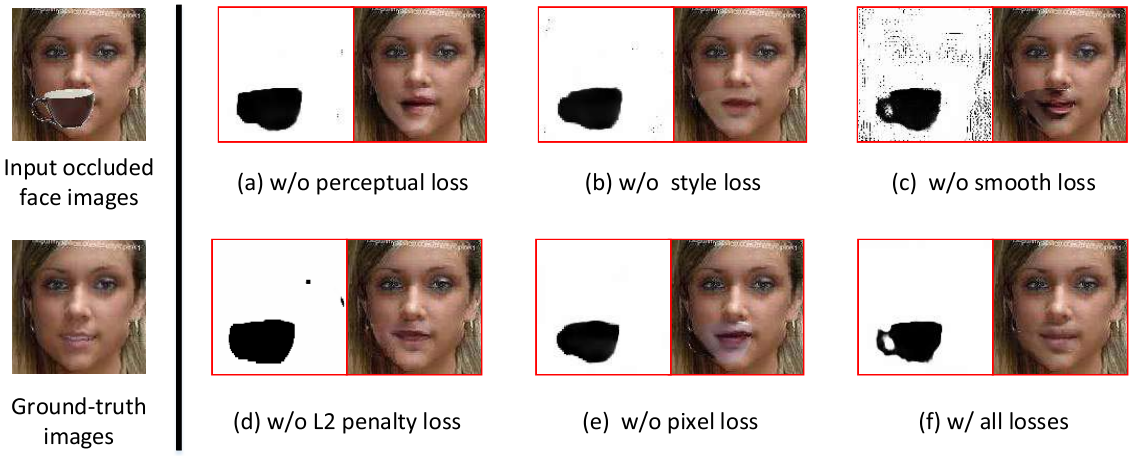


Fig. 14. Visualization of face de-occlusion results by OA-GAN in ablation study w.r.t. the joint losses: (a) without using perceptual loss, (b) without using style loss, (c) without using smooth loss, (d) without using pixel loss, (e) without using L_2 penalty loss, and (f) using all losses.

ground-truth face image. In addition, we observe that SSIM scores of individual methods are more consistent to the face recognition performance.

V. Ablation Study

Our OA-GAN uses different joint losses to optimize the network during training, so it is reasonable to verify the effectiveness of each loss function. We conduct ablation study of all the loss items (perceptual loss, style loss, pixel loss, smooth loss, and L_2 penalty loss) similar to [1], [13]. We discard one loss item from the loss function each time, and give the qualitative comparison of occlusion prediction and de-occlusion results in Figure 14. We can notice that discarding any loss item from our loss function may lead to artifacts in the face de-occlusion results. We also provided the PSNR and SSIM scores for individual methods during ablation study in Table IV. Again, we notice that each item in our loss function contributes to the convergence of our model.

We also study the benefit of our training strategy for the model convergence. We conducted two experiments: (i) pre-train with the paired face images of artificial occlusions and then finetune the model with the unpaired face images with natural occlusions and (ii) train with paired face images of artificial occlusions and unpaired face images of natural occlusion using our alternating training method. Since we do not have the ground-truth for face images with natural occlusion, we provide qualitative comparison about the recovered face images. As shown in Figure 15, we perceive that our alternating training leads to more visually pleasing face de-occlusion results. We believe that this benefits from the training with artificial occlusions. However, pre-training in experiment (i) does not have such an effect. Besides, if the model is trained directly without our strategy, we cannot obtain reasonable face de-occlusion results.

We also compare our two-stage generator with the existing two-path generator in [22], [23] (shown in Figure 4). For fair comparisons, we replace the generator of OA-GAN with the two-path generator in [22], [23], and then use our training strategy to train the network. The results are shown in Figure 16. We can see that compared with the two-path generator, the proposed two-stage generator can make better use of the contextual information to obtain better image de-occlusion results.

TABLE IV
PSNR AND SSIM OF THE DE-OCCLUSION RESULTS BY OUR METHOD WHEN DISCARDING INDIVIDUAL LOSS TERMS FOR ABLATION STUDY

Ablation condition	(a) w/o perceptual loss	(b) w/o style loss	(c) w/o smooth loss	(d) w/o L_2 penalty loss	(e) w/o pixel loss	(f) w/ all losses
PSNR(dB)	22.48	21.91	21.74	22.55	21.74	**22.61**
SSIM	0.784	0.726	0.760	0.781	0.742	**0.787**

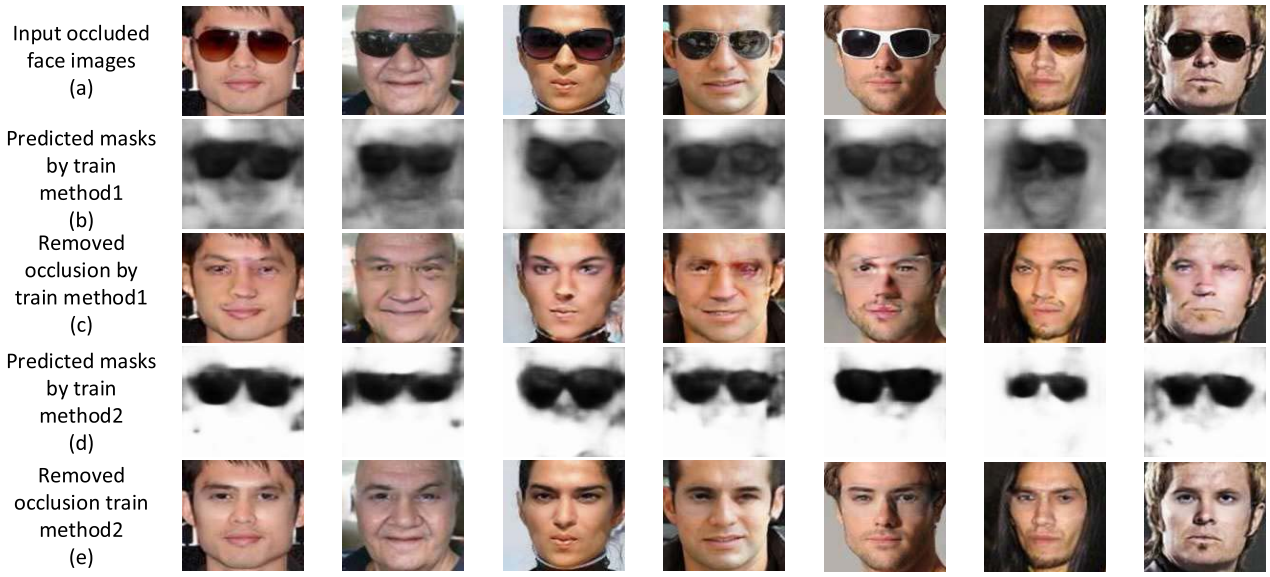


Fig. 15. Quantitative comparisons of different training strategies: firstly pre-training in paired face images with artificial occlusion and then fine-tuning in unpaired face images with natural occlusion (training method 1), and our proposed alternating training method (training method 2): (a) the input occluded face images, (b) and (d) the predicted masks using two different training methods, and (c) and (e) the de-occluded face images using two different training methods.

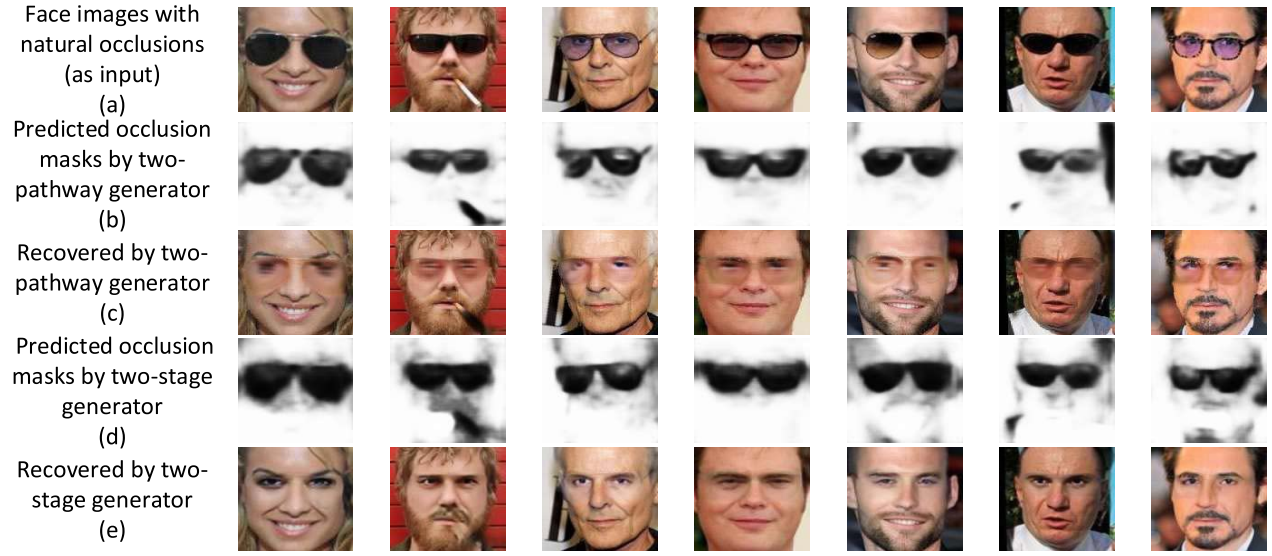


Fig. 16. Qualitative comparisons between the two-path generator and our two-stage generator: (a) the input images, (b) and (c) the predicted occlusion masks and the recovered face images by OA-GAN using the two-path generator, and (d) and (e) the predicted occlusion masks and the recovered face images by OA-GAN using the two-stage generator.

VI. CONCLUSION

In this paper, we propose a weakly supervised deep generative adversarial network (named as OA-GAN) for natural face de-occlusion. The proposed OA-GAN learns from unpaired face images with natural occlusion and paired face images with artificial occlusion in a semi-supervised manner using different joint loss functions. In addition, we design an alternate training strategy for model learning, which leads to better network convergence. Experimental results on the public CelebA dataset and a dataset with natural occlusions show

that the proposed approach can achieve promising results in recovering face images with unknown natural occlusions, and is helpful for improving face recognition performance.

In our future work, we would like to investigate new architecture designs of the generator and discriminator to recover high-quality face images from occluded face images. We also would like to leverage 3D face modeling, e.g., [40], [41] to improve face de-occlusion performance.

REFERENCES

[1] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE CVPR*, Jul. 2017, pp. 3911–3919.

[2] B.-W. Hwang and S.-W. Lee, "Reconstruction of partially damaged face images based on a morphable face model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 3, pp. 365–372, Mar. 2003.

[3] S. Zhang, R. He, Z. Sun, and T. Tan, "DeMeshNet: Blind face inpainting for deep MeshFace verification," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 637–647, Mar. 2018.

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

[5] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. CVPR*, Jun. 2018, pp. 5505–5514.

[6] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Trans. Image Process.*, vol. 10, no. 8, pp. 1200–1211, Aug. 2001.

[7] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 882–889, Aug. 2003.

[8] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. CGIT*, 2000, pp. 417–424.

[9] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[10] A. Telea, "An image inpainting technique based on the fast marching method," *J. Graph. Tools*, vol. 9, no. 1, pp. 23–34, Jan. 2004.

[11] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," in *Proc. SIGGRAPH*, 2009, p. 24.

[12] M.-C. Sagong, Y.-G. Shin, S.-W. Kim, S. Park, and S.-J. Ko, "PEPSI: Fast image inpainting with parallel decoding network," in *Proc. IEEE CVPR*, Jun. 2019, pp. 11360–11368.

[13] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," 2018, *arXiv:1804.07723*. [Online]. Available: http://arxiv.org/abs/1804.07723

[14] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *Proc. IEEE CVPR*, Jun. 2019, pp. 1486–1494.

[15] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. ICCV*, Dec. 2015, pp. 3730–3738.

[16] J. Cai, H. Han, S. Shan, and X. Chen, "FCSR-GAN: Joint face completion and super-resolution via multi-task learning," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 2, pp. 109–121, Apr. 2020.

[17] A. M. Martínez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 6, pp. 748–762, Jun. 2002.

[18] H. J. Oh, K. M. Lee, and S. U. Lee, "Occlusion invariant face recognition using selective local non-negative matrix factorization basis images," *Image Vis. Comput.*, vol. 26, no. 11, pp. 1515–1523, Nov. 2008.

[19] R. Min, A. Hadid, and J.-L. Dugelay, "Improving the recognition of faces occluded by facial accessories," in *Proc. FG*, Mar. 2011, pp. 442–447.

[20] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, Oct. 2017, pp. 2242–2251.

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

[22] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "GANimation: Anatomically-aware facial animation from a single image," in *Proc. ECCV*, 2018, pp. 818–833.

[23] C. Yang, T. Kim, R. Wang, H. Peng, and C.-C.-J. Kuo, "Show, attend, and translate: Unsupervised image translation with self-regularization and attention," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4845–4856, Oct. 2019.

[24] L. Xu, H. Zhang, J. Raitoharju, and M. Gabbouj, "Unsupervised facial image de-occlusion with optimized deep generative models," in *Proc. IEEE IPTA*, Nov. 2018, pp. 1–6.

[25] H. Han, J. Li, A. K. Jain, S. Shan, and X. Chen, "Tattoo image search at scale: Joint detection and compact representation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2333–2348, Oct. 2019.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.

[27] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022

[28] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, Jul. 2017, pp. 5967–5976.

[29] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NeurIPS*, 2014, pp. 2672–2680.

[30] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015, *arXiv:1508.06576*. [Online]. Available: http://arxiv.org/abs/1508.06576

[31] H. Zhang, J. Yang, Y. Zhang, and T. S. Huang, "Non-local kernel regression for image and video restoration," in *Proc. ECCV*, 2010, pp. 566–579.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[33] S. Ge, J. Li, Q. Ye, and Z. Luo, "Detecting masked faces in the wild with LLE-CNNs," in *Proc. CVPR*, Jul. 2017, pp. 2682–2690.

[34] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein GANs," 2017, *arXiv:1704.00028*. [Online]. Available: http://arxiv.org/abs/1704.00028

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[36] 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.

[37] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," 2019, *arXiv:1905.12384*. [Online]. Available: http://arxiv.org/abs/1905.12384

[38] J. Cui, H. Zhang, H. Han, S. Shan, and X. Chen, "Improving 2D face recognition via discriminative face depth estimation," in *Proc. IEEE ICB*, Feb. 2018, pp. 140–147.

[39] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

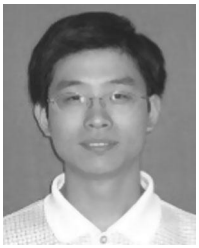
[40] H. Han and A. K. Jain, "3D face texture modeling from uncalibrated frontal and profile images," in *Proc. BTAS*, Sep. 2012, pp. 223–230.

[41] K. Niinuma, H. Han, and A. K. Jain, "Automatic multi-view face recognition via 3D model based pose regularization," in *Proc. BTAS*, Sep. 2013, pp. 1–8.

Jiancheng Cai received the B.S. degree from Shandong University in 2017, and the M.S. degree from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), and the University of Chinese Academy of Sciences in 2020. He is currently a Research Engineer with Meituan-Dianping. His research interests include computer vision, pattern recognition, and image processing with applications to biometrics.

Hu Han (Member, IEEE) received the B.S. degree from Shandong University, and the Ph.D. degree from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), in 2005 and 2011, respectively, both in computer science. He is currently an Associate Professor with ICT, CAS. Before joining the faculty at ICT, CAS, in 2015, he has been a Research Associate at the PRIP Laboratory, Department of Computer Science and Engineering, Michigan State University, and a Visiting Researcher at Google, Mountain View, CA, USA. His research interests include computer vision, pattern recognition, and image processing, with applications to biometrics and medical image analysis. He has authored or coauthored over 60 papers in refereed journals and conferences, including the IEEE TPAMI, TIP, TIFS, TBIOM, CVPR, ECCV, NeurIPS, and MICCAI. He was a recipient of the IEEE FG2019 Best Poster Award, and CCBR 2016/2018 Best Student/Poster Awards. He is an Associate Editor of *Pattern Recognition*, an Area Chair of ICPR2020.

Jiyun Cui received the B.S. degree from Northeast Normal University, in 2016, and the master's degree from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), in 2019. He is currently a Research Engineer at Baidu. His research interests include computer vision and pattern recognition, with the focus on bio-perception oriented intelligent computing.



Jie Chen (Member, IEEE) received the M.S. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 2002 and 2007, respectively. He joined the faculty with the Graduate School in Shenzhen, Peking University, in 2019, where he is currently an Associate Professor with the School of Electronics and Computer Engineering. Since 2018, he has been working with the Peng Cheng Laboratory, Shenzhen, China. From 2007 to 2018, he worked as a Senior Researcher with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. In 2012 and 2015, he visited the Computer Vision Laboratory, University of Maryland, and the School of Electrical and Computer Engineering, Duke University, respectively. His research interests include deep learning, computer vision, and medical image analysis. He was the Co-Chair of International Workshops at ACCV, CVPR, ICCV, and ECCV. He was a Guest Editor of special issues for IEEE TPAMI, IJCV, and *Neurocomputing*. He is an Associate Editor of the *Visual Computer*.

Li Liu (Senior Member, IEEE) received the B.Sc. degree in communication engineering, the M.Sc. degree in photogrammetry and remote sensing, and the Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), China, in 2003, 2005, and 2012, respectively. She is currently an Associate Professor with the College of System Engineering. During her Ph.D. study, she has spent more than two years as a Visiting Student at the University of Waterloo, Canada, from 2008 to 2010. From 2015 to 2016, she spent ten months visiting the Multimedia Laboratory, The Chinese University of Hong Kong. From 2016 to 2018, she worked as a Senior Researcher at the Machine Vision Group, University of Oulu, Finland. She was the Co-Chair of nine International Workshops at CVPR, ICCV, and ECCV. She serves as the Guest Editor of special issues for IEEE TPAMI and IJCV. Her current research interests include computer vision and machine learning. Her papers have currently about 3000 citations in Google Scholar. She serves as the Area Chair for ICME2020 and ACCV2020. She currently serves as an Associate Editor for *Pattern Recognition Letters* and the *Visual Computer Journal*.

S. Kevin Zhou (Fellow, IEEE) received the Ph.D. degree from the University of Maryland, College Park, MA, USA. He is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences. Prior to this, he was a Principal Expert and a Senior Research and Development Director of Siemens Healthcare. He has published over 200 book chapters and peer-reviewed journal and conference papers, registered more than 130 granted patents, written two research monographs, and edited three books. His three most recent books are *Medical Image Recognition, Segmentation and Parsing: Machine Learning and Multiple Object Approaches*, SK Zhou (Ed.), *Deep Learning for Medical Image Analysis*, SK Zhou, H Greenspan, DG Shen (Eds.)," and *Handbook of Medical Image Computing and Computer-Assisted Intervention*, SK Zhou, D Rueckert, G Fichtinger (Eds.) He has been elected as a board member of the MICCAI Society, an Advisory Board Member of MONAI (Medical Open Network for AI), and a fellow of the AIMBE. He has won multiple awards including the R&D 100 Award (Oscar of Invention), Siemens Inventor of the Year, and UMD ECE Distinguished Alumni Award. He has been the Program Co-Chair for MICCAI2020, Lima, Peru, an Associate Editor for the IEEE TRANSACTIONS MEDICAL IMAGING AND MEDICAL IMAGE ANALYSIS, and an Area Chair for CVPR, MICCAI, NeurIPS, and AAAI.