IEEE Transactions on Information Forensics & Security

Semi-supervised Natural Face De-occlusion

Journal:	*Transactions on Information Forensics & Security*
Manuscript ID	T-IFS-11015-2020.R1
Manuscript Type:	Regular Paper
Date Submitted by the Author:	08-Jun-2020
Complete List of Authors:	Cai, Jiancheng; Institute of Computing Technology Chinese Academy of Sciences, Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences; Han, Hu; Institute of Computing Technology Chinese Academy of Sciences, ; Cui, Jiyun; Institute of Computing Technology Chinese Academy of Sciences Chen, Jie; Peng cheng laboratory Liu, Li; National University of Defense Technology, College of System Engineering; Oulun Yliopisto, Center for Machine Vision and Signal Analysis Zhou, Shaohua Kevin; Siemens Corporate Research,
EDICS:	BIO-MODA-FAC-Face biometrics < BIO-BIOMETRICS, SUR-OTHS-Others < SUR-SURVEILLANCE

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*

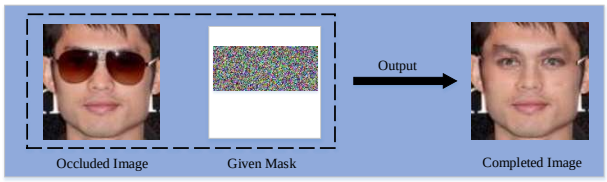
***Abstract*—Occlusions are often present in face images in the wild, e.g., under video surveillance and forensic scenarios. Existing face de-occlusion methods are limited as they require the knowledge of an occlusion mask. To overcome this limitation, we propose in this paper a new generative adversarial 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.**

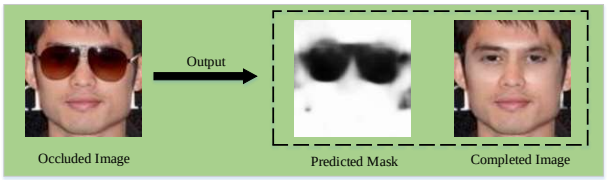
I. Introduction

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 solve the occluded scenarios with little 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.



(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)

Figure 1. (a) Existing methods require a given mask, and filling the mask with Gaussian noise. (b) Our method jointly predicts the occlusion area and recovers the image. While existing methods, e.g., GFC [1], require a given mask for face de-occlusion, the proposed approach can jointly predict the occlusion regions and recover the image.

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.

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.

Jiancheng Cai, Hu Han, Jiyun Cui, 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. E-mail: {jiancheng.cai, jiyun.cui}@vipl.ict.ac.cn; {hanhu, zhoushaohua}@ict.ac.cn

Jie Chen is with the Peking University Shenzhen Graduate School. E-mail: chenj@pcl.ac.cn

Li Liu is with the University of Oulu, Finland and the National University of Defense Technology, China. Email: li.liu@oulu.fi

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

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

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

(iv) The proposed approach outperforms the state-of-the-art 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. [15] 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 [16] 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. [17] 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 [17] 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 (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 [18] divided the face images into k local regions and designed a probabilistic method to analyze the occlusions in each region. JunOh et al. [19] also divided the face image into a finite number of

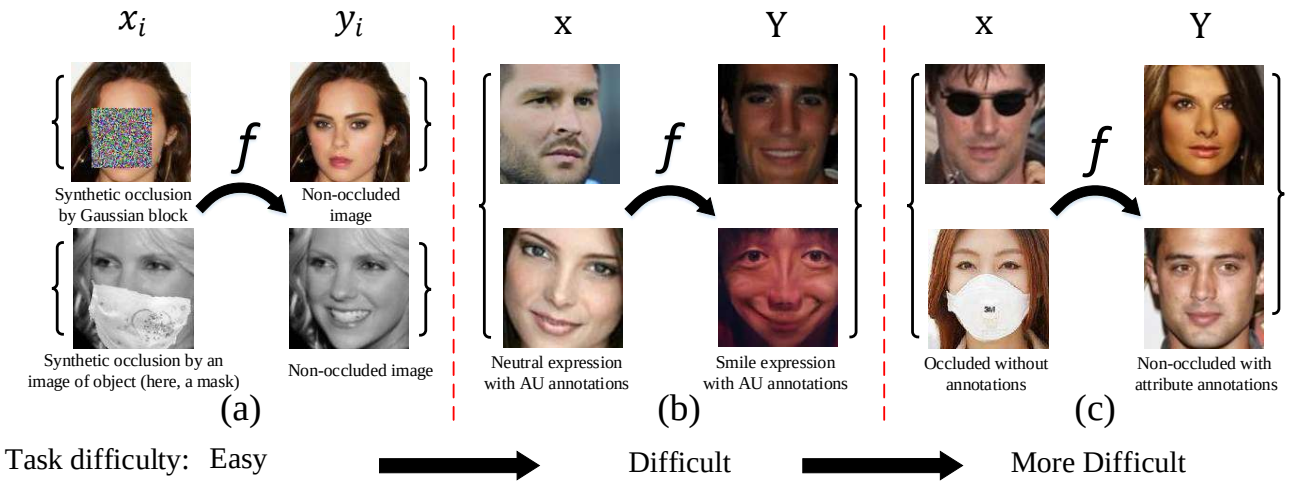


Figure 2. (a) Paired training data $\{x_i, y_i\}_i$, in which x_i is the face image with artificial occlusion and y_i is its corresponding non-occluded face image. (b) Unpaired training data $\{X, Y\}$ used in the expression conversion task, in which X consists of face images with neutral expression 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.

disjoint local patches and then determined whether each patch belongs to the occluded area or not using a PCA model. Min et al. [20] 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. Li et al. [21] employed a PG-Unit to tell if every divided facial region occluded or not and then reweighted the local facial regions for better facial expression recognition.

The face occlusion detection in our OA-GAN is different 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 [22] and Dual-GAN [23] 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 [24] and Self-regularization [25] used the attention mechanism so that the network can accurately locate the facial regions to be modified and achieved promising face attribute conversion results. He et al. [26] proposed an AttGAN using attribute classification constraint to edit facial attributes. However, this method can only edit face attributes in a small area, and it is difficult to complete the face image with large occlusion.

In this paper, we focus on recovering non-occluded face images from face images with natural occlusions. Similar to the face attribute conversion, we also use a semi-supervised approach to build our face recovery network. In the network structure, proposed method is similar to the GANimation [24] and Self-regularization [25], both can predict the mask and generate image. While GANimation and Self-regularization use two-pathway generator, proposed method uses two-stage generator. We will discuss the difference of the two generator in the next Section. In addition, compared to these face attribute conversion methods, we introduce image comletion loss aiming for better face completion.

III. Proposed Approach

Our goal of this work is to learn a face de-occlusion model in a semi-supervised fashion. Specifically, we aim to tackle a more challenging face de-occlusion problem that deals with real world occlusions without the knowledge of 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 with natural occlusion x_i in X, there is NO ground-truth face image y_i in Y that corresponds to x_i. 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 image de-occlusion, where paired face images are available for training. It is also more challenging

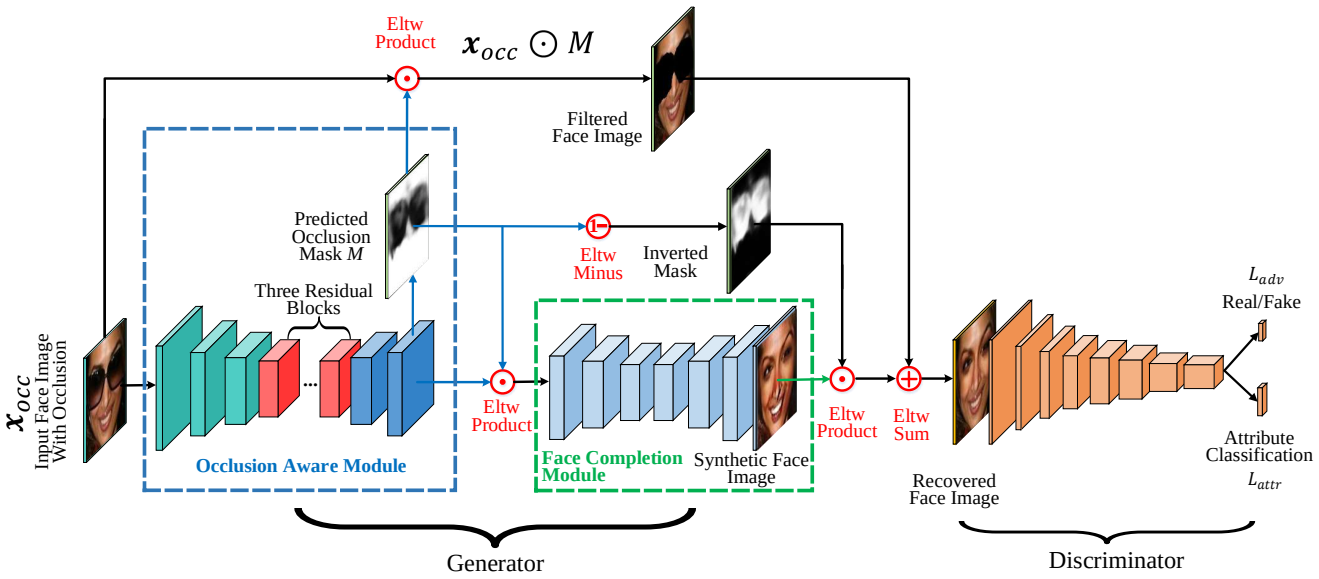


Figure 3. Overview of our occlusion-aware GAN (OA-GAN) for semi-supervised face de-occlusion without using either natural paired training images or occlusion masks.

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

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

than face expression conversion, which is actually an image style change. The reasons why natural face de-occlusion is more challenging are as follows. (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 [24] 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; (iii) While the expression conversion method [24] takes the cycle structure to 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 [24] (as shown in Figure 2 (b)). For [24], 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, the expression conversion method is difficult than the existing face de-occlusion method, and our face de-occlusion method is more difficult than the expression conversion method (as shown in Figure 2).

To this end, we propose an OA-GAN (as shown in 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. 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 the occlusions, respectively. The face occlusion-aware module has an encoder-decoder architecture consisting of six residual blocks [27], with instance normalization [28] and ReLU layers after every convolution and deconvolution layers in our generator, and uses 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:

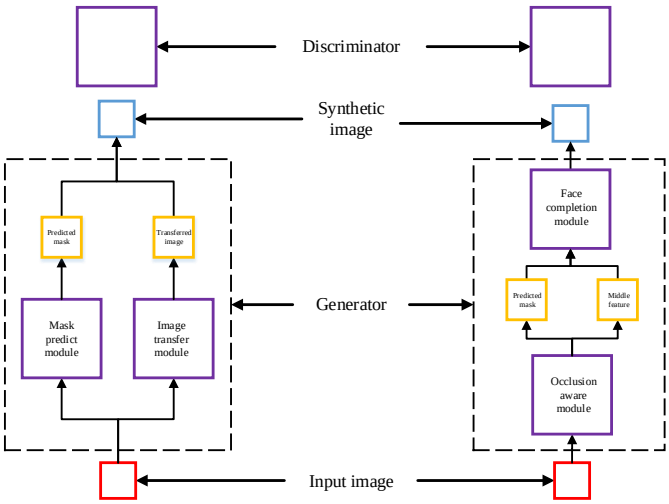
$$x_{final} = M \odot x \ + (1 - M) \odot x_{synth}, \tag{2}$$

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 in our approach is essentially different from the generators using in existing methods like GANimation [24] and Self-regularization [25]. GANimation [24] and Self-regularization [25] 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, particularly the face completion module, can leverage the contextual information from the occlusion-aware module (see Figure 4 (b)) to achieve better face completion results. Specifically, the first stage of the generator (occlusion prediction) will be optimized based on no 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 the original input face image. In our experiments, the supervisory signal of



(a) GANimation and Self-regularization (b) OA-GAN

Figure 4. Different network structures between (a) the generator used in GANimation [24] and Self-regularization [25] and (b) the generator of our OA-GAN.

attributions can offset the influence of unbalanced data. For example, in the CelebA dataset, there are significantly less senior people than young people, 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 PatchGAN [29], but with modifications of the last layer adding an attribute classifier. The detailed discriminator architecture 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 the L_{adv} is the adversarial loss [30] and the L_{attr} is the mean square error of the attribute between the ground-truth image and the recovered face image. Here, we only compute the L_{attr} for synthetic paired face images. α and β are two hyper-parameters that balance the influences of the two losses.

C. Alternating Training

The absence of natural paired face images (with and without natural occlusions) poses additional challenges to the face de-occlusion task. To make the network training possible, we propose an alternating training strategy, with different loss combinations in different stages, to optimize the whole network. Overall, the alternating training consists of two stages: (i) auxiliary training with synthetic paired images, and (ii) training with natural unpaired images (as shown in Figure 5). The former gives the network a good ability to complete the face images and the later aims to ensure the position of the occlusion.

Auxiliary training with synthetic paired images. In this stage, for synthetic paired images, we combine multiple loss functions as defined below to train the network.

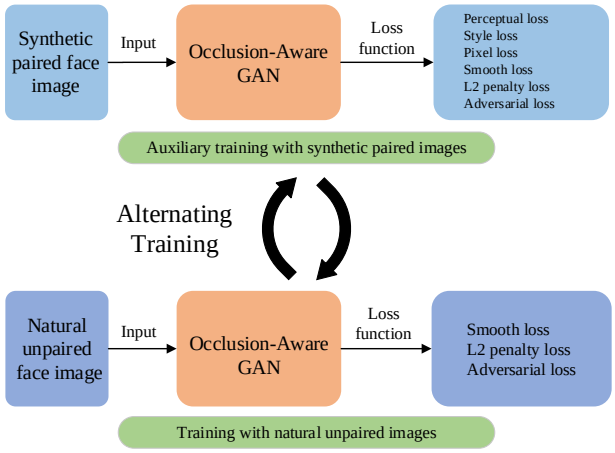


Figure 5. The diagram of alternating training for our OA-GAN.

The perceptual loss [31] 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:

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

where the ϕ is the VGG-16 [32] which is pre-trained based on ImageNet, and ϕ_n represents the $n\text{-}th$ 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:

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

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

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

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

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

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

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

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:

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

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

Training with natural unpaired images. In this stage, for natural unpaired face images, we introduce the smoothness loss, the mask L_2 penalty loss and the adversarial loss to train the network. The total loss function for the natural unpaired face image is defined as follow:

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

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 [17] in that: while two-stage training is a step-by-step optimization for different modules of the network using the same data, 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 natural unpaired face images becomes 1 to 1.

IV. Experimental Results

A. Database

We perform experimental evaluations on the public CelebA dataset [16], which consists of 202,599 face images from 10,177 subjects, with each image annotated with 40 binary attributes. We treat glasses as 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 19746 face images from CelebA, the MAFA [33] and the Internet. In our experiments, randomly select 80% of the naturally occluded face

Figure 6. Face occlusion-aware and face completion results by our proposed approach for face images with artificial occlusion. (a) shows the input images with artificial occlusion. (b) shows the masks predicted by our proposed method. (c) and (d) show the de-occluded and ground-truth face images.

images (by either glasses or glasses) and non-occluded face dataset for training, and the remaining 20% face images for testing.

We build the paired occluded and non-occluded face dataset using more than 640 different types of artificial occlusions (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 scaled to 128×128. For the face images we collected, we locate the five facial landmarks using an open-source landmark detection algorithm SeetaFace[1]. Although occlusion 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 even the face landmark detection accuracy is not good for face images with occlusion. During training, our artificial occlusions are generated by randomly placing an object on the face image as shown in Figure 6 (a).

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 [13], [1] 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 the hyper-parameters ($\lambda_5 = -1$, $\lambda_6 = 1$) by making the scale of the value the same as the scale of the value of the first four items. Although we only use such a simple rule to choose the hyper-parameters, the face de-occlusion visualization and face recognition results (see Figure 7 and Figure 13) show the effectiveness of the proposed approach.

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

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 on face occluded image dataset. The first experiment is to verify the effectiveness of the proposed OA-GAN in recovering from face images with artificial occlusion (see Figure 6). We can see that the proposed approach generates visually reasonable results compared with the ground-truth face images. The important facial structures and characteristics also look visually similar to the ground-truth.

The second experiment is to compare the results of our OA-GAN in recovering from face images with natural occlusions with Cycle-GAN [22], Self-regularization [25], and GANimation [24], which are used for converting natural unpaired images between two domains or face expression conversion. The results are shown in Figure 7. In our experiments, Cycle-GAN, [25], and GANimation only recover the face images from one type of occlusion; so two models need to be trained for recovering face images from occluded face images with glasses and respirators, respectively. From Figure 7, we can see the results of our method are much better than those obtained by Cycle-GAN, Self-regularization, and GANimation. We consider that Cycle-GAN and GANimation are based on the cycle structure which can convert an image (I_a) from domain A to domain B (denoted as I_{ab}), and then convert image I_{ab} back to domain A during the model training. Such a cycle structure can leverage self-supervision to perform image transformation between two domains, and has been widely used in tasks like face attribute conversion [24]. Face de-occlusion is a more

Figure 7. Qualitative comparisons of face de-occlusion results by Cycle-GAN [22], GANimation [24], Self-regularization [25] 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.

complicated task than style transformation, which requires that the non-occluded facial area must remain the same as that in the input occluded face image, while the occluded area can be reasonably recovered. The cycle structure cannot assure such a requirement during transformation between two domains. Besides, although Self-regularization does not rely on a cycle structure, and uses a generator with two-path structure, it does not make good use of the contextual information during face completion.

In the third experiment, we compare the proposed OA-GAN with [36], which can perform natural face de-occlusion using DCGANs in an iterative way. Our approach differs from [36] in the following aspects: (i) [36] tries to generate a non-occluded face image from a control vector. In the generated face image, the area outside the occlusion mask are expected to be as similar as the same area of the input face image. Different from [36], the second-stage of our generator (the face completion module) generates a non-occluded face image by using the features of the area outside the occlusion mask, which is the output of the first-stage of our generator (the occlusion-aware module). As a result, our method can generate a non-occluded face image with higher quality than [36]. (ii) [36] is an iterative method; even during network inference, it usually requires about 1000 iterations, which requires much more computational cost than our single forward inference pass. (iii) While [36] does not require any paired data (e.g., occluded face image and mated non-occluded face image) for training, our method can leverage the knowledge from

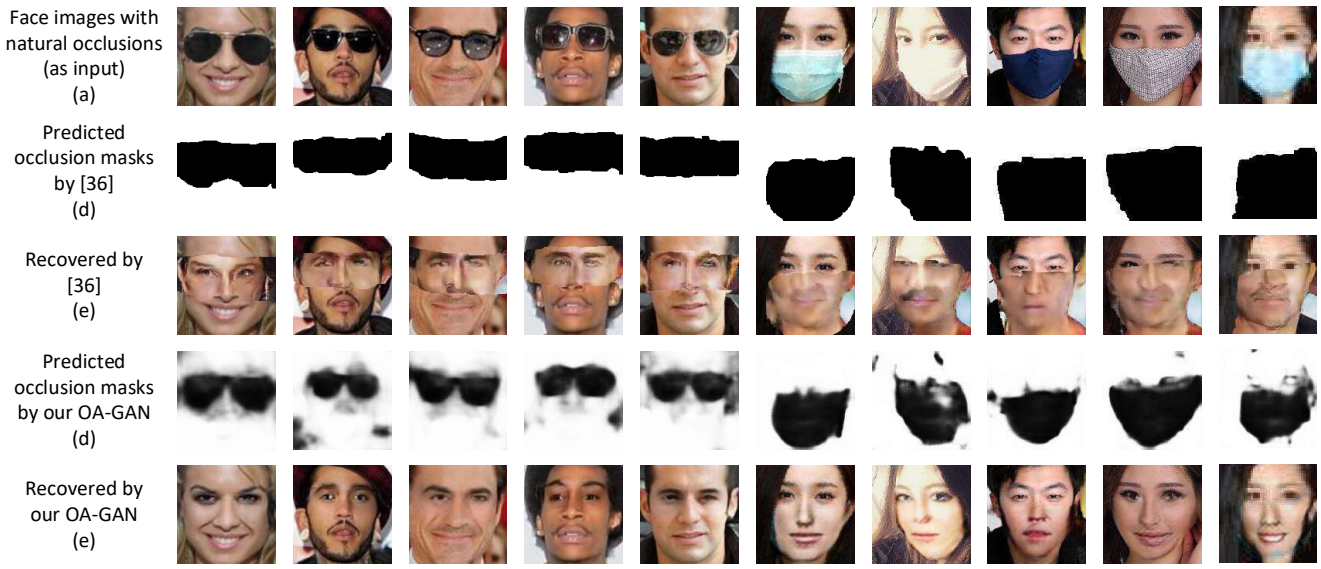


Figure 8. Qualitative comparisons of face de-occlusion results by [36] and our OA-GAN for several face images with natural occlusion from the CelebA database. (a) are the original input face images. (b, d) are the predicted occlusion masks by [36] and our OA-GAN, respectively. (c, e) are the recovered face images by and our OA-GAN, respectively.

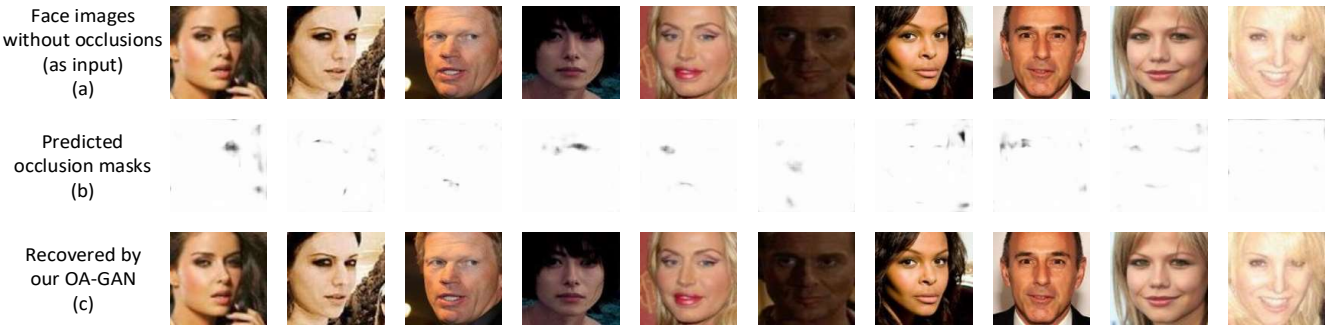


Figure 9. Results of the proposed OA-GAN in dealing with face images without occlusion. (a) the input face images, (b) the predicted occlusion masks, and (c) the face images after de-occlusion.

paired face images with synthetic occlusion to perform semi-supervised model learning for natural face de-occlusion. Visual comparisons of the face de-occlusion results by our method and [36][2] in Figure 8 show the superiority our approach.

In the four experiment, 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 few 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. For de-occlusion experiments using face images with natural occlusions, it is difficult to find the exact ground-truth face images without occlusions.

D. Quantitative Comparisons

In addition to visual quality, we conducted two different experiments to quantify the effectiveness of the proposed approach. In the first experiment, we use two metrics to evaluate proposed method on recovering synthetic face images. One metric is the peak signal-to-noise ratio (PSNR), which is widely used in image compression area to measure the fidelity of the reconstructed image. The other metric is the structural similarity index (SSIM) [37], which 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 report the results in terms of PSNR and SSIM on the synthetic occluded face image dataset.

We compare the proposed approach with face completion method GFC [1], and general image inpainting methods GnIpt [5], Pconv [13] and CSA [38]. For fair comparisons, we use the 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 occlusion mask as input, we use the mask predicted by our model, together with the occluded face image, 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

[2]Since the code of [36] is not publicly available, we reimplemented the method in [36] based on the best of our understanding.

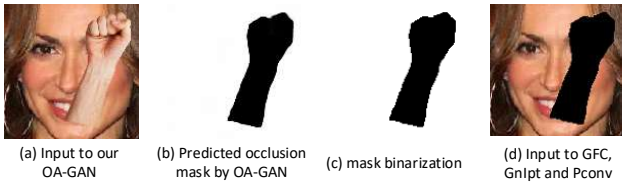


Figure 10. (a) the input face image with artificial occlusion, (b) predicted occlusion mask by our OA-GAN, (c) is the binary mask image for (b), and (d) the input to GFC, GnIpt, Pconv and CSA, which is the multiplication result of (a) and (c).

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 [38]) 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

to recover the face images using our mask. The PSNR and SSIM achieved by the proposed approach, GFC, GnIpt, Pconv and CSA[3] 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 artificial occlusions. These results suggest that while the existing methods on face de-occlusion may not work when manual occlusion masks are not available, the proposed OA-GAN can still obtain very reasonable de-occlusion results.

In the second experiment, we performed user study by asking three participants to select the best one from the face de-occlusion results by our OA-GAN, and three state-of-the-art (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. 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 by our method can have better perceptual quality than the SOTA methods.

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 choose LightCNN-9 [39] as a face recognition model and use two types of face images to train it: (a) original face images with natural occlusions in CelebA, and (b) the face images after de-occlusion by OA-GAN. The training, gallery, and probe sets for face identification contain

[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.

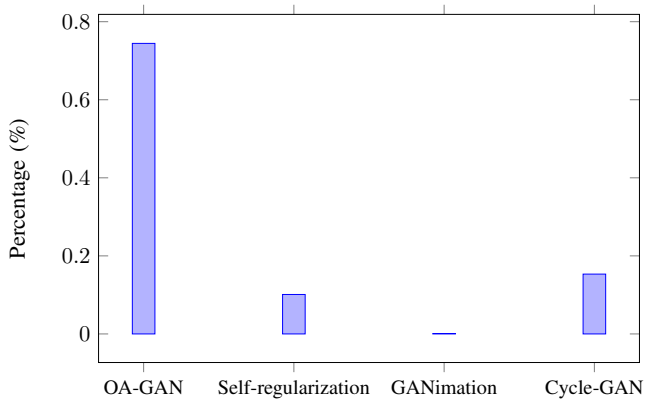


Figure 11. The percentage of best face de-occlusion results in user study for our OA-GAN, GANimation, Cycle-GAN and self-regularization.

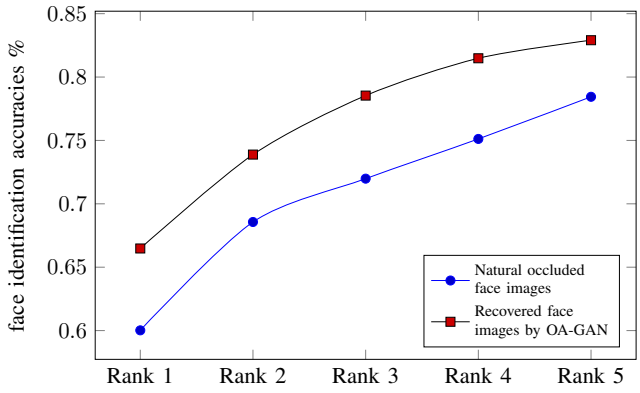


Figure 12. Rank 1-5 face identification accuracies using natural occluded face images and recovered face images by OA-GAN.

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 occluded face images. We also visualized 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 occlusion, 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 ground-truth image.

V. ABLATION STUDY

Our proposed OA-GAN uses different losses to optimize the network when training with the unpaired face images with natural occlusions and paired face images with artificial occlusions. In order to verify the effect of each loss function, we conduct ablation study about all the loss items (perceptual loss, style loss, pixel loss, smooth loss, and L2 penalty loss) in our loss function referring to existing methods [13], [1]. We discard one loss item from the loss function each time,

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

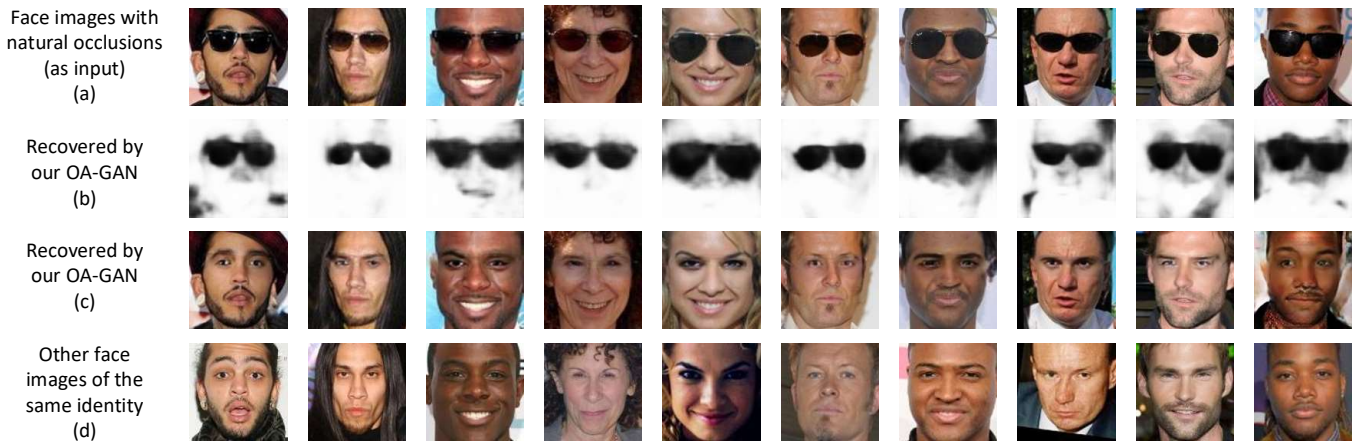


Figure 13. Comparison between the recovered face images by OA-GAN and another face image of the same person. (a) are the input face images. (b) are the predicted occlusion masks. (c) are the recovered face images by our OA-GAN, and (d) are another face images of the same person.

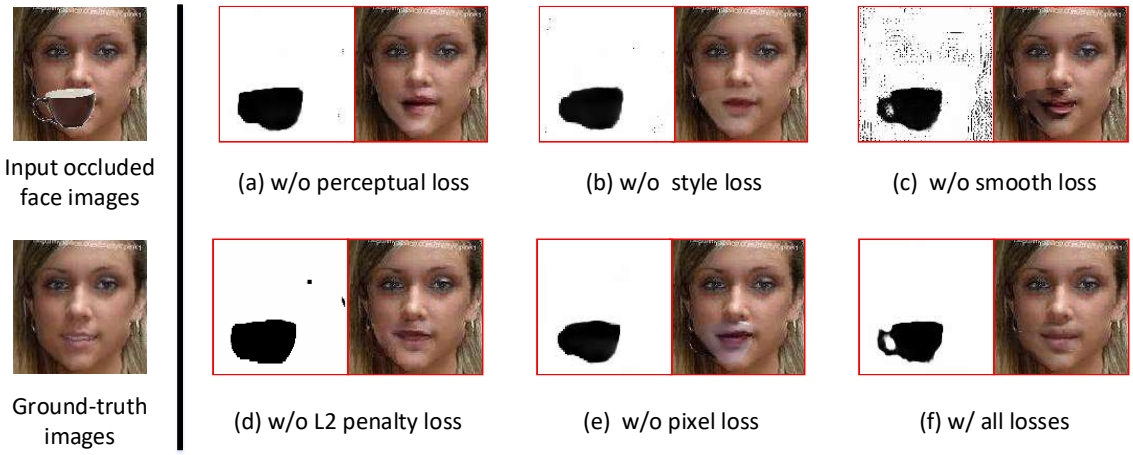


Figure 14. Visualization of face de-occlusion results by OA-GAN. (a) are the input face images with natural occlusion, (b) are the predicted occlusion masks, (c) are the recovered face images by our OA-GAN, and (d) are another face image of the same person, which are not the exact ground-truth face images, but can be used as references of the ground-truth.

Table IV
PSNR AND SSIM OF THE DE-OCCLUSION RESULTS BY OUR METHOD DURING 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**

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 severe artifacts in the face images after de-occlusion. 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 facial images. 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.

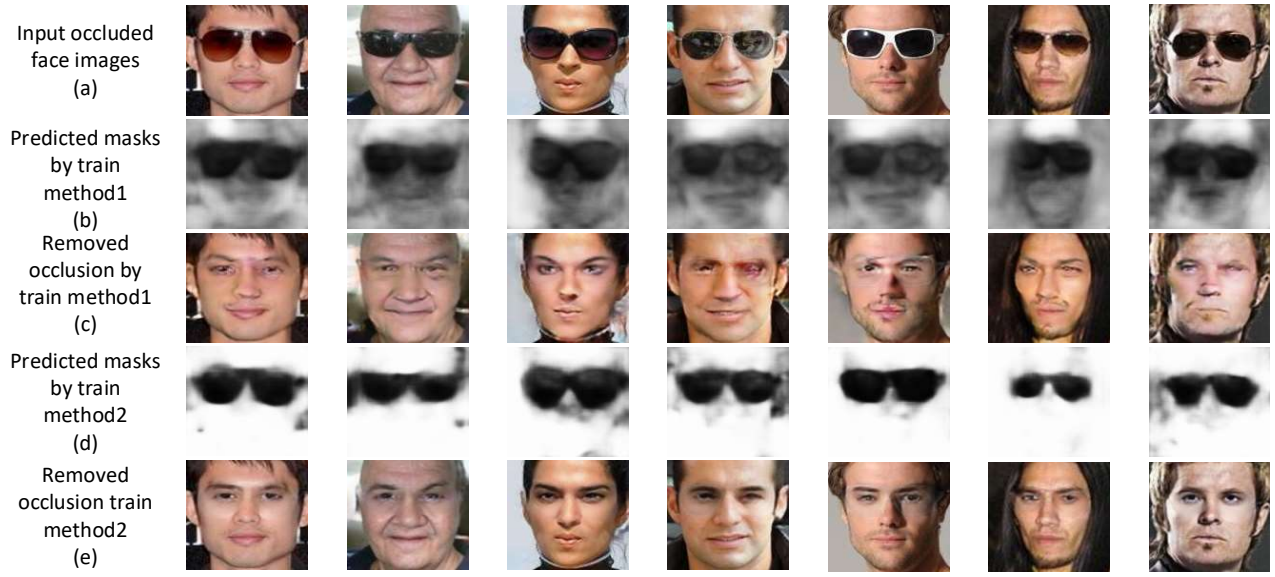


Figure 15. Quantitative comparisons of firstly pre-training in paired face images with artificial occlusion and then fine-tuning in unpaired face images with natural occlusion (training method 1), and our proposed alternating training method (training method 2). (a) is the input of the occluded face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

We also compare our two-stage generator with the existing two-path generator[24], [25] (shown in Figure 4). For fair comparisons, we replace the generator of OA-GAN with the two-path generator, 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 network structure can make better use of contextual information to obtain reasonable image de-occlusion results.

VI. Conclusion

In this paper, we propose a 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 loss functions. The smoothness loss, the L_2 weight penalty loss, and adversarial loss are used for the natural unpaired occluded face image. For the synthetic paired images, apart from above three losses, the perceptual loss, style loss, and pixel loss are added. Besides, we design an alternate training strategy to obtain 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 designs of the generator and discriminator to recover high-quality face images from occluded face images. We also would like to study the face de-occlusion method utilizing 3D face priors [40][41] and apply the recovered face images for face recognition, attribute learning, face parsing, emotion recognition, etc.

Acknowledgment

This research was supported in part by the Natural Science Foundation of China (grants 61732004 and 61672496), External Cooperation Program of Chinese Academy of Sciences (CAS) (grant GJHZ1843), and Youth Innovation Promotion Association CAS (2018135).

References

[1] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. CVPR*, 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, 2003.
[3] S. Zhang, R. He, and T. Tan, "Demeshnet: Blind face inpainting for deep meshface verification," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 637–647, 2018.
[4] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. CVPR*, 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*, 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, 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, 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, 2004.
[10] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of graphics tools*, vol. 9, no. 1, pp. 23–34, 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.
[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*, 2019, pp. 11 360–11 368.

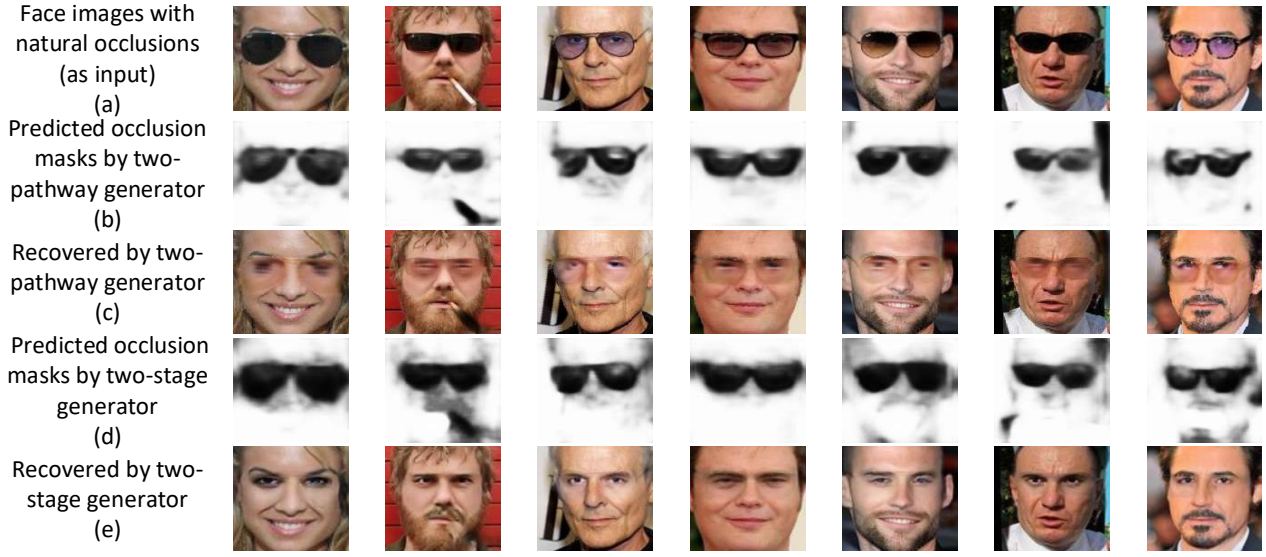


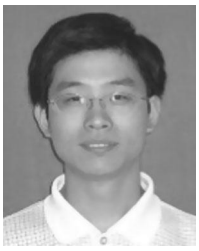
Figure 16. Qualitative comparisons between the two-path generator and our two-stage generator. (a) is the input images, (b) and (c) are the predicted occlusion masks and the recovered face images by OA-GAN using the two-path generator, (d) and (e) are the predicted occlusion masks and the recovered face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

[13] G. Liu, F. A. Reda, K. J. Shih, and T.-C. Wang, "Image inpainting for irregular holes using partial convolutions," *arXiv preprint, arXiv:1804.07723*, 2018.
[14] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *Proc. IEEE CVPR*, 2019, pp. 1486–1494.
[15] 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, 2018.
[16] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. ICCV*, 2015.
[17] J. Cai, H. Han, S. Shan, and X. Chen, "Fcsr-gan: Joint face completion and super-resolution via multi-task learning," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2019.
[18] 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, 2002.
[19] H. JunOh, K. MuLee, and S. UkLee, "Occlusion invariant face recognition using selective local non-negative matrix factorization basis images," *Image and Vision computing*, vol. 26, no. 11, pp. 1515–1523, 2008.
[20] R. Min, A. Hadid, and J.-L. Dugelay, "Improving the recognition of faces occluded by facial accessories," in *Proc. FG*, 2011, pp. 442–447.
[21] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-gated cnn for occlusion-aware facial expression recognition," in *Proc. ICPR*, 2018.
[22] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017, pp. 2242–2251.
[23] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation." in *Proc. ICCV*, 2017, pp. 2849–2857.
[24] 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.
[25] 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, 2019.
[26] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE Trans. on Image Process.*, 2019.
[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.
[28] D. Ulyanov and A. Vedaldi, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.
[29] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017, pp. 5967–5976.
[30] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NeurIPS*, 2014, pp. 2672–2680.
[31] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.
[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[33] S. Ge, J. Li, Q. Ye, and Z. Luo, "Detecting masked faces in the wild with lle-cnns," in *Proc. CVPR*, 2017, pp. 2682–2690.
[34] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," *arXiv preprint arXiv:1704.00028*, 2017.
[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[36] L. Xu, H. Zhang, J. Raitoharju, and M. Gabbouj, "Unsupervised facial image de-occlusion with optimized deep generative models," pp. 1–6, 2018.
[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
[38] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," *arXiv preprint arXiv:1905.12384*, 2019.
[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, 2018.
[40] H. Han and A. K. Jain, "3D face texture modeling from uncalibrated frontal and profile images," in *Proc. BTAS*, 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*, 2013, pp. 1–8.

Jiancheng Cai received the B.S. degree from Shandong University in 2017, and he is working toward the M.S. degree in the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), and the University of Chinese Academy of Sciences. His research interests include computer vision, pattern recognition, and image processing with applications to biometrics.

Hu Han is an Associate Professor of the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS). He received the B.S. degree from Shandong University, and the Ph.D. degree from ICT, CAS, in 2005 and 2011, respectively, both in computer science. Before joining the faculty at ICT, CAS in 2015, he has been a Research Associate at PRIP lab in the Department of Computer Science and Engineering at Michigan State University, and a Visiting Researcher at Google in Mountain View. His research interests include computer vision, pattern recognition, and image processing, with applications to biometrics and medical image analysis. He has authored or co-authored over 50 papers in refereed journals and conferences including IEEE Trans. PAMI/IP/IFS/BIOM, 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 a member of the IEEE.

Jiyun Cui received the BS degree from Northeast Normal University, in 2016, and the master degree from ICT, CAS in 2019. He is now a research engineer at Baidu. His research interests include computer vision and pattern recognition with focus on bio-perception oriented intelligent computing.



Jie Chen received the MSc and PhD degrees from the Harbin Institute of Technology, 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 Electronic and Computer Engineering. Since 2018, he has been working with the Peng Cheng Laboratory, 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 School of Electrical and Computer Engineering, Duke University respectively. His research interests include pattern recognition, computer vision, machine learning, deep learning, and medical image analysis. He is an associate editor of the Visual Computer. He is a member of the IEEE.

Li Liu (SM'19) received the BSc, MSc and PhD degrees from the National University of Defense Technology (NUDT), China, in 2003, 2005, and 2012, respectively. She joined the faculty with NUDT in 2012, where she is currently an associate professor with the College of System Engineering. From 2008.1 to 2010.3, she visited the University of Waterloo, Canada. From 2015.3 to 2016.1, she visited the Multimedia Laboratory, Chinese University of Hong Kong. From 2016.11 to 2018.11, she worked with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. She was a cochair of seven International Workshops at CVPR, ICCV and ECCV. She was the leading guest editor of special issues of the IEEE Transactions on Pattern Analysis and Machine Intelligence and the nternational Journal of Computer Vision. Her current research interests include facial behavior analysis, image and video descriptors, object detection, and recognition. Her papers have about 2,000 citations in Google Scholar. She is a senior member of the IEEE.

S. Kevin Zhou Professor S. Kevin Zhou obtained his PhD degree from University of Maryland, College Park. He is a Professor at Chinese Academy of Sciences. Prior to this, he was a Principal Expert and a Senior R&D director at Siemens Healthcare. Dr. Zhou has published 180+ book chapters and peer-reviewed journal and conference papers, registered 250+ patents and inventions, written two research monographs, and edited three books. His two most recent books are entitled "Medical Image Recognition, Segmentation and Parsing: Machine Learning and Multiple Object Approaches, SK Zhou (Ed.)" and "Deep Learning for Medical Image Analysis, SK Zhou, H Greenspan, DG Shen (Eds.)." He has won multiple awards including R&D 100 Award (Oscar of Invention), Siemens Inventor of the Year, and UMD ECE Distinguished Aluminum Award. He has been an associate editor for IEEE Transactions on Medical Imaging and Medical Image Analysis, an area chair for CVPR and MICCAI, a board member of the MICCAI Society. He is a Fellow of IEEE and AIMBE.