Look More Into Occlusion: Realistic Face Frontalization and Recognition With BoostGAN

Qingyan Duan, *Student Member, IEEE*, and Lei Zhang, *Senior Member, IEEE*

***Abstract*—Many factors can affect face recognition, such as occlusion, pose, aging, and illumination. First and foremost are occlusion and large-pose problems, which may even lead to more than 10% accuracy degradation. Recently, generative adversarial net (GAN) and its variants have been proved to be effective in processing pose and occlusion. For the former, pose-invariant feature representation and face frontalization based on GAN models have been studied to solve the pose variation problem. For the latter, frontal face completion on occlusions based on GAN models have also been presented, which is much concerned with facial structure and realistic pixel details rather than identity preservation. However, synthesizing and recognizing the occluded but profile faces is still an understudied problem. Therefore, in this article, to address this problem, we contribute an efficient but effective solution on how to synthesize and recognize faces with large-pose variations and simultaneously corrupted regions (e.g., nose and eyes). Specifically, we propose a boosting GAN (BoostGAN) for occluded but profile face frontalization, deocclusion, and recognition, which has two aspects: 1) with the assumption that face occlusion is incomplete and partial, multiple images with patch occlusion are fed into our model for knowledge boosting, i.e., identity and texture information and 2) a new aggregation structure integrated with a deep encoder–decoder network for coarse face synthesis and a boosting network for fine face generation is carefully designed. Exhaustive experiments on benchmark data sets with regular and irregular occlusions demonstrate that the proposed model not only shows clear photorealistic images but also presents powerful recognition performance over state-of-the-art GAN models for occlusive but profile face recognition in both the controlled and uncontrolled environments. To the best of our knowledge, this article proposes to solve face synthesis and recognition under poses and occlusions for the first time.**

***Index Terms*—Face frontalization, face recognition, face synthesis, generative adversarial net (GAN).**

I. Introduction

FACE recognition has recently achieved a great progress, enabled by convolutional neural network-based deep learning techniques [1]–[5]. However, there are many problems still remained to be unresolved satisfactorily. The most important aspect is the large-pose variation, which is a general bottleneck of face recognition. The existing methods that address kinds of pose variations can be generally divided into two fundamental categories. Specifically, one category of methods aims to obtain pose-invariant features of multiview facial images by using handcrafted descriptors or deep learning models [3], [6]. Traditional feature representation approaches for pose-invariant face recognition include HOG [6], SIFT [7], LBP [8], Gabor [9], and well-known sparse representation [10], [11]. However, these handcrafted feature descriptors cannot well extract the identity discriminative feature; as a result, face recognition performance is limited. The appearance of deep learning [12] has tackled the bottleneck. A series of convolutional networks has been developed to recognize faces and close the gap between human and machine [1], [13], [14]. Moreover, the face recognition performance of deep learning-based approaches has already significantly exceeded human ability on some benchmarks [3]–[5], [15]–[19].

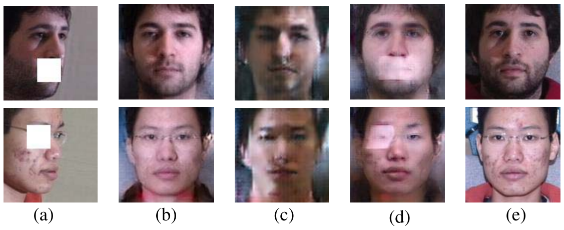

Fig. 1. Synthesis results by testing the existing models on occluded faces. The poses of the first row and the second row are 45° and 60°, respectively. GT denotes the ground-truth frontal images. The comparisons of face frontalization by removing occlusion first are also presented in Section VI. (a) Profile. (b) Ours. (c) [23]. (d) [24]. (e) GT.

Beyond the deep representation-based face recognition, very recently, inspired by the generative adversarial network (GAN) technique [20], a number of GAN variant-based synthesis approaches have been proposed to generate photorealistic frontal view facial images [21]–[28]. The generator in GANs often is deployed with an encoder–decoder based on convolutional neural network architecture, in which the encoder extracts identity preserved features, while the decoder outputs photorealistic faces with a target pose controlled by the pose code [23], landmark heatmaps [25], appearance flow [26], and so on. In addition to pose variation, illumination was also considered as an important factor in the pose-invariant feature representation, synthesis, and recognition [23], [29].

Manuscript received April 23, 2019; revised September 20, 2019, December 19, 2019, and January 5, 2020; accepted February 29, 2020. Date of publication March 19, 2020; date of current version January 5, 2021. This work was supported in part by the National Science Fund of China under Grant 61771079, in part by the Chongqing Natural Science Fund under Grant cstc2018jcyjAX0250, and in part by the Chongqing Youth Talent Program. *(Corresponding author: Lei Zhang.)*

The authors are with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China (e-mail: qyduan@cqu.edu.cn; leizhang@cqu.edu.cn.)

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

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

Digital Object Identifier 10.1109/TNNLS.2020.2978127

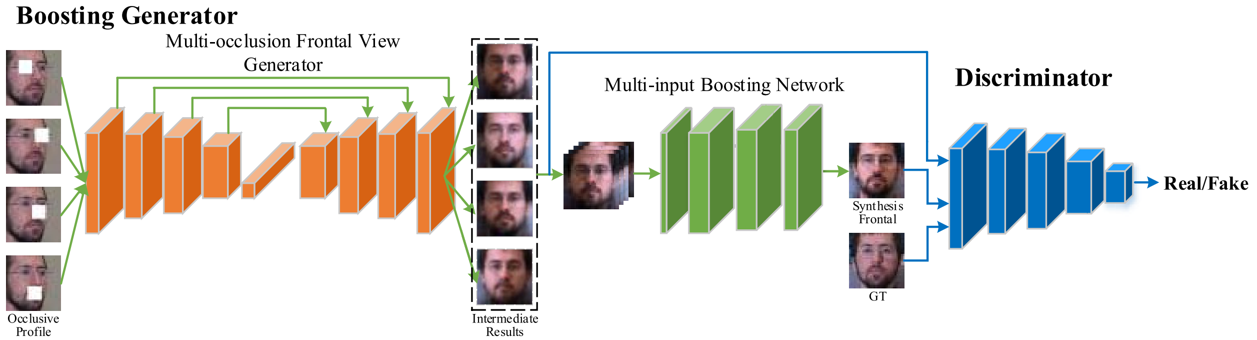

Fig. 2. Framework of the BoostGAN with an end-to-end coarse-to-fine architecture. It includes two stages, i.e., multiocclusion frontal-view generator and multi-input boosting network. In the first stage, the coarse frontal faces with deocclusion and slight identity preservation are generated. In the second stage, the photorealistic frontal faces without occlusions are eventually synthesized by fusing the information from multiple inputs in a boosting ensemble way.

In fact, not only the variation of poses but also occlusions seriously affect the face recognition performance. In order to fill the "hole" of the faces, it is natural to consider the image completion methods. Most of the traditional approaches based on low-level cues tend to synthesize the missing contents in the "hole" by searching similar patches from other regions of the same image [30]–[32]. Recently, deep learning [33]–[36], especially GAN model, has also been presented for the task of image completion [37]–[40]. However, these methods often compute and infer the pixels in the corrupted image region while without changing the pixels of the uncorrupted regions. In addition, the most important objective of image completion is to maintain the consistency of detail and texture such that the semantic facial structure can be restored, rather than preserving identity information. This means that these methods are more useful to the task of close-set image completion, rather than the open-set task. This is because, in the open-set image completion task, it is impossible to find the matched clean image in the training stage for the corrupted pixels of the test images. Consequently, the filled pixels of the test images are imprecise and lack of identity discrimination, that is, if a facial image with occlusion is excluded from the training set, then the repaired face cannot preserve the raw identity, which is clearly not conducive to the recognition task of faces. Although Zhao *et al.* [41] introduced an identity preservation term in the training stage of the face completion network, it is not sufficient to faultlessly solve the open-set problem.

There are many methods proposed to solve the pose and occlusion issues separately. However, in practice, both two variations often exist simultaneously on faces, which becomes a challenging but understudied issue in face synthesis and recognition. In this article, to address this challenge, we contribute to answer how to recognize faces if occlusion and large-pose variation coexist simultaneously? Actually, this is a well-known problem of face recognition, and our solution to this problem is clearly the boosting GAN (BoostGAN) model proposed in this article. To the best of our knowledge, this is the first work for integrating the face deocclusion and frontalization tasks together based on the GAN model with the goal of face recognition. The existing methods of face frontalization often frontalize the clean profile faces without considering occlusions. However, once the face is occluded or corrupted in some region, the effect of synthesis of these previous methods will become very poor, as clearly described in Fig. 1. In addition, the existing image completion approaches were designed for restoring the occlusion part of a near-frontal view face, rather than profile faces. Although the "hole" of an image can be filled by using GAN-based image completion approaches (i.e., deocclusion), the identity and texture information cannot be preserved in the synthesized faces, which is not beneficial to face recognition.

With the previous face completion and frontalization models [23], [24], [37]–[40], it is natural to wonder that two-step processing, i.e., face completion for deocclusion as the first step and face frontalization as the second step, may be employed to solve this problem. However, the current face completion methods are unable to process profile images and simultaneously preserving the identity information such that the unfaithful deoccluded results in the first step will affect the generated frontal facial image in the second step and further affect the recognition accuracy. In addition, the two-step processing will incur a large number of network parameters and depend on a mass of training samples with complicated and expensive data collection and annotation.

Motivated by the abovementioned problems, in this article, an end-to-end BoostGAN model is proposed to solve the face frontalization and recognition challenge when both occlusions and pose variations exist simultaneously. The structure of BoostGAN is presented in Fig. 2, which is first reported in our reprint version [42]. A novel coarse-to-fine aggregation architecture deployed with a deep encoder–decoder network (i.e., coarse net) and a shallow boosting network (i.e., fine net) is presented for identity-invariant photorealistic frontal-view face synthesis. For network training, four different occluded profile images from the same person are prepared as an input so that the complementary information can be extracted and utilized to generate multiple frontal-view faces by the coarse net. Then, the fine net employs the multiple intermediate outputs of the coarse net to further generate a photorealistic and identity-preserved frontal face, that is, our BoostGAN adopts a progressive generation strategy, by first generating multiple intermediate coarse faces (deocclusion) and second synthesizing photorealistic frontal face. Different from the previous GAN-based face completion, both the pixel- and the feature-level information from multiple scales are modeled as the supervisory signal in order to preserve the face identity information. Thus, BoostGAN is guaranteed to deal with the open-set occluded but profile face synthesis and recognition.

Compared with the existing work [23], [24], [37], [38], the main new contributions in this article lie in threefolds.

1) An end-to-end BoostGAN model is proposed for recognizing occlusive but profile faces, in which photorealistic but identity-preserved frontal view faces can be synthesized from arbitrary occlusions and poses. Our end-to-end model greatly reduces the amount of parameters and training complexity compared with two-step method. To the best of our knowledge, this is the first work for addressing the coexistence of pose and occlusion.
2) A coarse-to-fine aggregation structure with progressive generation strategy is presented. The coarse part is a deep encoder–decoder network that is used for deocclusion and frontalization of multiple profile faces with partial occlusion. Multiple intermediate faces with deocclusion can be generated by the coarse net. The fine part is deployed with a shallow boosting network for photorealistic face frontalization with identity preservation.
3) Three types of occlusions, such as regular keypoint region occlusion, regular square random occlusion, and irregular occlusion on profile images of different poses, are tested in our experiments. Quantitative and qualitative experiments on benchmark data sets validate the superiority of the proposed BoostGAN over state-of-the-art (SOTA) methods under both constrained and unconstrained scenarios.

II. Related Work

In this section, three closely related topics, including generative adversarial network (GAN), face frontalization, and image completion, are reviewed.

A. Generative Adversarial Network

GAN, proposed by Goodfellow *et al.* [20], was presented for image generation through gaming between a generator and a discriminator. Because of its unprecedented performance, GAN and its variants have been exploited in various tasks of visual synthesis, such as image-to-image translation [43], [44], image super-resolution [45]–[47], style transfer [48]–[50], and face attribute manipulation [51]–[53]. Recently, several typical GAN models were introduced. Deep convolutional GAN (DCGAN) [54] was the first work to integrate the convolutional network into GAN for better photorealistic image generation. Then, many different GAN variants are proposed. However, the training stability of GAN models is always an open problem, which therefore has attracted many researchers to solve the training instability problem of GANs. For example, Wasserstein GAN [55] removed the logarithm function in the original GAN losses. Spectral normalization generative adversarial network (SNGAN) [56] proposed the spectral normalization in order to satisfy the Lipschitz constraint and guarantee the statistics bound. Besides, there are many other works focusing on the improvement of visual realism. For example, Zhu *et al.* [57] proposed the cycleGAN to deal with the unpaired data. Karras *et al.* [58] generated the high-resolution image from a low-resolution image by progressively improving both the generator and the discriminator. DA-GAN [59] combined prior knowledge and domain knowledge of faces to exactly and inherently recover the information by projection from 3-D to 2-D space.

B. Face Frontalization

Due to its ill-posed nature, face frontalization is an extremely challenging task. The existing methods that cope with this ill-posed problem can be generally divided into three categories: 2-D/3-D local texture warping [21], [60], statistical methods [22], and deep learning methods [23]–[25], [61]. To be specific, Hassner *et al.* [21] proposed to exploit a mean reference surface of the 3-D face to compute a face image of frontal view for each subject. Sagonas *et al.* [22] recognized the frontal-view reconstruction and landmark localization as a constrained low-rank minimization problem. Inspired by GAN, Tran *et al.* [23] proposed a DR-GAN model for pose-invariant feature representation and recognition, which aims to extract pose-invariant feature representation from the encoder. FF-GAN [62] integrated the 3-D face model into GAN model such that the 3-D morphable model (3DMM) conditioned model can keep the visual quality during face frontalization. TP-GAN [24], as a typical and successful model, proposed to cope with the profile faces by using global and local networks separately and synthesize the final frontal-view face using network fusion in order to improve the photorealism of generated images. CAPG-GAN [25] recovers the neutral and profile head pose images from the input face and landmark heatmap with pose-guided generator and discriminator. Zhang *et al.* [26] introduced spatial transformer networks (STNs) [63] to this task, and then, the dense correspondences between the profile and frontal faces can be established. In this way, the synthetic frontal-view faces can preserve the detailed facial textures. PIM [64] proposed to unify the subnet of face frontalization (FFN) and the subnet of discriminative learning (DLN) in an end-to-end deep architecture for pose-invariant recognition. The 3-D PIM [65] unified a simulator for 3-D face reconstruction and frontal-view synthesis, and a refiner for realism refinement addressed the challenging large-pose variations. HF-PIM [66] frontalized profile images with higher resolution using a novel facial texture fusion warping, which combines the advantages of 3-D model and GAN. UV-GAN [67] proposed to fit a 3DMM and then employed local and global networks to learn identity-preserved UV completion.

These methods focus on the impact of pose variation and work well under nonocclusive scenarios. However, the effect of occlusion is not considered. Due to the specificity of these network architectures, the existing methods cannot be generalized to synthesize frontal-view faces when pose and occlusion coexist, as described in Table I.

C. Image Completion

Filling the missing pixels of a facial image can be recognized to be image completion. Content prior that usually comes from the uncorrupted parts of the same image or an external data set is often required in order to obtain

TABLE I

FEASIBILITY OF GAN VARIANTS FOR PROFILE FACE FRONTALIZATION AND RECOGNITION WITH/WITHOUT OCCLUSION

Method	Without occlusion	With occlusion
DR-GAN [23]	✓	✗
FF-GAN [62]	✓	✗
TP-GAN [24]	✓	✗
CAPG-GAN [25]	✓	✗
BoostGAN	✓	✓

a confident restoration. The early algorithms tend to inpaint the missing pixels by propagating the information from the known neighborhoods with low-level cues or global statistics, that is, they search the similar structures from the context of the input image and then place them into the holes [30]–[32]. Deep neural network-based method tried to restore the missing content in the image from the background texture [68]. Recently, GANs have been proposed for the same task [38]–[40]. Specifically, Li *et al.* [38] proposed a GAN-based image completion model with global and local discriminators. Yeh *et al.* [69] computed the missing pixels and regions by using the semantic image inpainting-based data. However, they are incapable of preserving the facial identity. Therefore, Zhao *et al.* [41] presented to recover the missing pixel parts under different head poses while trying to preserve the identity based on an identity loss and a pose discriminator in network training.

These methods of image completion only compute and fill the missing pixels and work on the frontal or near-frontal occluded face for image inpainting. In addition, the information of identity cannot be well preserved with these image completion methods. Therefore, recognizing faces under the coexistence of both occlusion and pose variations is still a challenge and understudied.

III. PROPOSED BOOSTGAN MODEL

Different from those methods of face frontalization and image completion, our proposed BoostGAN model can well work on the challenging scenario that pose variations and occlusions coexist, for synthesizing and recognizing photorealistic frontal-view faces.

A. Network Architecture

The proposed BoostGAN network consists of three parts: 1) multiocclusion frontal-view generator; 2) multi-input boosting network; and 3) multi-input discriminator. The details of each part are presented in the following.

1) Multiocclusion Frontal View Generator: In this article, two primary kinds of regular block occlusions with respect to different positions are considered: keypoint position occlusion and random position occlusion, as shown in Fig. 2 (the inputs). There are four occlusive profile images $I^{b_i}, i \in \{1, 2, 3, 4\}$, which are occluded in the positions of the left eye, the right eye, the nose, and the mouth on a profile face by a white block mask, respectively. This is recognized as keypoint position occlusion. Notably, the frontal ground truth of a profile image is represented as I^{gt}. The size of I^{b_i}, $i \in \{1, 2, 3, 4\}$, and I^{gt} is $W \times H \times C$, where W, H, and C mean width, height, and channel number of input, respectively. The objective of the multiocclusion frontal face generator is to generate four roughly frontal images from the four profile images with different occlusions. The multiocclusion frontal face generator G^c consists of an encoder and decoder, represented as $\{G_e^c$ and $G_d^c\}$, and works as a coarse generator, that is

$$I^{f_i} = G^c(I^{b_i}), \quad i \in \{1, 2, 3, 4\} \tag{1}$$

where $I^{f_i}, i \in \{1, 2, 3, 4\}$, is the roughly generated frontal-view face corresponding to each different occluded profile facial image. It is worth noting that, when only one occlusive profile image is prepared, our proposed method is also feasible, by simply replicating multiple images. Inspired by TP-GAN [24], the generator G^c is formulated with an encoder of downsampling and a decoder of upsampling with skip connections.

2) Multi-Input Boosting Network: As mentioned earlier, four rough and slightly discriminative frontal faces can be coarsely obtained by the multiocclusion frontal-view image generator G^c. However, due to the keypoint region that is occluded, the identity preservation capability is not good and the photorealistic quality of synthesis is flawed with blur and distortion. Therefore, to improve the photorealism and identity preservation that better beneficial to face recognition, a multi-input boosting network is then constructed and followed after the coarse generator G^c for photorealistic and identity-preserved face generation. By boosting the complementary information of multiple inputs, the image quality is improved, which is called a fine generator. Boosting is an ensemble strategy-based metaalgorithm, which is capable of transforming a group of weak learners into a strong learner by model fusion. Therefore, it is natural to convert the four primarily generated frontal images from the coarse generator into a photorealistic frontal face with better identity information by using the proposed boosting network (fine generator).

Specifically, the boosting network that used to deal with high-quality image generation is represented as G^f. The four primarily generated frontal images from G^c are concatenated to feed as input (the size is $W \times H \times 4C$) of the proposed boosting network, which is shown as follows:

$$I^f = G^f(I^{f_1}, I^{f_2}, I^{f_3}, I^{f_4}) \tag{2}$$

where I^f represents the generated photorealistic and identity preserved frontal image. The proposed boosting network contains two residual blocks, and each block is deployed with a convolutional layer. Leaky-ReLU and batch normalization are used so as to avoid overfitting and gradient vanishing. The configuration of the boosting network is shown in Table II.

3) Multi-Input Discriminator: Different from most of GANs, the input of the discriminator of BoostGAN is not an image pair (i.e., real image versus synthetic image) but multiple inputs. Specifically, in BoostGAN, the input of multiocclusion frontal-view generator includes four occlusive profile images and then four coarse frontal faces after deocclusion (i.e., $I^{f_i}, i \in \{1, 2, 3, 4\}$) can be obtained. For the boosting network, these four primarily generated images are fed as the input, and the finally generated photorealistic and identity-preserved frontal face I^f can be acquired. So far, totally, five

TABLE II

CONFIGURATION OF THE BOOSTING NETWORK. THE RESBLOCK DENOTES THE RESIDUAL CONNECTION

Layer	Input	Filter Size	Output Size
resblock1	Images	$\begin{bmatrix} 5\times 5, 12 \\ 5\times 5, 12 \end{bmatrix} \times 1$	$128\times 128\times 12$
conv1	resblock1	$5\times 5, 64$	$128\times 128\times 64$
resblock2	conv1	$\begin{bmatrix} 3\times 3, 64 \\ 3\times 3, 64 \end{bmatrix} \times 1$	$128\times 128\times 64$
conv2	resblock2	$3\times 3, 32$	$128\times 128\times 32$
conv3	conv2	$3\times 3, 3$	$128\times 128\times 3$

synthetic images are synthesized by the coarse and fine generators of BoostGAN. Therefore, six images (i.e., five synthetic images and one ground-truth frontal view image) are employed as the input of discriminator, which is therefore named multi-input discriminator. The multi-input discriminator is optimized as a binary classifier to distinguish real images and synthetic images. The network structure is similar to [24].

B. Training Loss Functions

The BoostGAN is trained from scratch end-to-end by taking the weighted sum of multiple losses as the supervisory signal, such as adversarial loss, pixel-level losses (i.e., pixel, symmetric, and total variation), and identity preserving loss.

1) Adversarial Loss: In the training stage, three components are coarse generator G^c, fine generator G^f, and discriminator D, where the inputs of G^c are profile faces with occlusions, the inputs of G^f are the coarsely synthesized outputs of G^c, and the goal of D is to distinguish the fake data generated by G^c and G^f from the real data. The generators G^c and G^f aim to synthesize photorealistic images to fool the discriminator D. The game between G (i.e., G^c and G^f) and D can be expressed as a min–max-based value function $V(D, G)$

$$\begin{aligned}\min_{G^c,G^f}\max_{D} V(D,G) &= \frac{1}{N}\sum_{n=1}^{N}\Bigg\{\log D(I_n^{gt}) + \frac{1}{5}\Bigg(\sum_{i=1}^{4}\log\big(1-D\big(G^c\big(I_n^{b_i}\big)\big)\big) \\ &\quad + \log\big(1-D\big(G^f\big(I_n^{f_1} I_n^{f_2}, I_n^{f_3}, I_n^{f_4}\big)\big)\big)\Bigg)\Bigg\} \qquad (3)\end{aligned}$$

where N is the batch size, I^{b_i}, $i = 1, \ldots, 4$, denotes the raw inputs with pose variations and occlusions, and I^{f_j}, $j = 1, \ldots, 4$, denotes the coarsely generated frontal-view faces by G^c. In practice, G^c, G^f, and D are obtained by alternatively solving the following optimization problems:

$$\begin{aligned}L_D &= \max_{D} V(D,G) \\ &= \frac{1}{N}\sum_{n=1}^{N}\Bigg\{\log D\big(I_n^{gt}\big) + \frac{1}{5}\Bigg(\sum_{i=1}^{4}\log\big(1-D\big(G^c\big(I_n^{b_i}\big)\big)\big) \\ &\quad + \log\big(1-D\big(G^f\big(I_n^{f_1} I_n^{f_2}, I_n^{f_3}, I_n^{f_4}\big)\big)\big)\Bigg)\Bigg\} \qquad (4)\end{aligned}$$

$$\begin{aligned}L_{\text{adv}} &= \max_{G^c,G^f} V(D,G) \\ &= \frac{1}{5N}\sum_{n=1}^{N}\Bigg\{\sum_{i=1}^{4}\log D\big(G^c\big(I_n^{b_i}\big) \\ &\quad + \log D\big(G^f\big(I_n^{f_1}, I_n^{f_2}, I_n^{f_3}, I_n^{f_4}\big)\big)\big)\Bigg\}. \qquad (5)\end{aligned}$$

2) Identity Preserving Loss: In the face deocclusion and face frontalization process, facial identity information is very easy to be corrupted, which is undoubtedly not beneficial to face recognition task. Therefore, in order to better preserve the human identity of the generated face image, the identitywise feature represented by a pretrained recognition network with face images is employed as supervisory perceptual signal. The perceptual loss formulated by an L_1-distance is exploited as the identity preserving loss function

$$L_{\text{ip}} = \sum_{i=1}^{5} |F^{po}(I^{gt}) - F^{po}(\hat{I}^i)| + |F^{fc}(I^{gt}) - F^{fc}(\hat{I}^i)| \qquad (6)$$

where $F^{po}(\cdot)$ and $F^{fc}(\cdot)$ represent the outputs of the last pooling layer and fully connected layer of the pretrained Light CNN [70] on large data sets of faces, respectively. $\hat{I}$ represents all the generated faces, including $\hat{I}^{f_j}$, $j \in \{1, 2, 3, 4\}$, and $\hat{I}^f$. Totally, five generated faces are computed.

3) Pixel-Level Losses: In order to further guarantee the content consistency of multi-image and improve the photorealism, three kinds of pixel-level losses, including multiscale pixelwise L_1 loss (pix), symmetry loss (sym), and total variation (tv) regularization [71], are exploited as follows:

$$L_{\text{pix}} = \sum_{i=1}^{5}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{W_k H_k C}\sum_{w,h,c=1}^{W_k,H_k,C}\big|\hat{I}^i_{k,w,h,c} - I^{gt}_{k,w,h,c}\big| \qquad (7)$$

$$L_{\text{sym}} = \sum_{i=1}^{5}\frac{1}{W/2\times H}\sum_{w=1}^{W/2}\sum_{h=1}^{H}\big|\hat{I}^i_{w,h} - \hat{I}^i_{w_s,h}\big| \qquad (8)$$

$$L_{\text{tv}} = \sum_{i=1}^{5}\sum_{c=1}^{C}\sum_{w,h=1}^{W,H}\big|\hat{I}^i_{w+1,h,c} - \hat{I}^i_{w,h,c}\big| + \big|\hat{I}^i_{w,h+1,c} - \hat{I}^i_{w,h,c}\big| \qquad (9)$$

where K represents the number of scales, W_k and H_k mean the width and height of scale k, respectively, and $w_s = W - (w - 1)$ is the symmetric abscissa of w in the generated faces $\hat{I}$. In the proposed BoostGAN model, three scales, including $32 \times 32, 64 \times 64$, and 128×128, are tested.

4) Overall Loss of Generator: In summary, the final loss function for training the generator G (i.e., G^c and G^f) is a weighted summation of five loss functions

$$L_G = \lambda_1 L_{\text{adv}} + \lambda_2 L_{\text{ip}} + \lambda_3 L_{\text{pix}} + \lambda_4 L_{\text{sym}} + \lambda_5 L_{\text{tv}} \qquad (10)$$

where λ_1–λ_5 are tradeoff parameters.

C. Two Different Testing Schemes

The main difference of our BoostGAN from other GAN variants is that the complementary information from four different occlusive inputs is ensembled and boosted. However, in

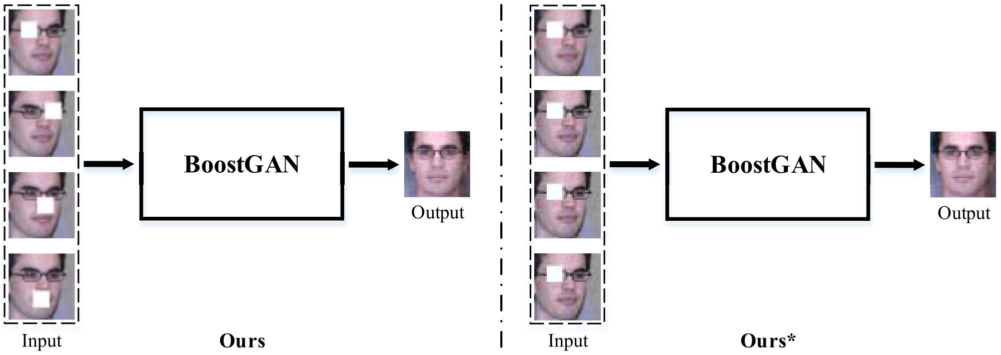

Fig. 3. Two different testing schemes with regard to inputs. Left: four faces with different keypoint occlusions are feeded as inputs, which is called Ours. Right: four faces with the same occlusions are feeded as inputs, which is called Ours*.

a real application, we cannot always obtain four different occlusive facial images. Therefore, to validate the effectiveness of our BoostGAN, two testing schemes with different strategies of inputs are considered, as shown in Fig. 3. First, as our model describes, four profile faces with different occlusions are used as an input, which is denoted as Ours. Second, when only one occlusive profile face image is prepared, we simply adopt the four copies of the only one occlusive image as inputs, which is represented as Ours* in experiments.

IV. Experiments Under Regular Occlusions

In this section, experiments are conducted under regular occlusions. To validate the superiority of our proposed BoostGAN for profile face frontalization and recognition with occlusions, qualitative and quantitative experiments on constrained and unconstrained benchmark data sets are employed. First, the qualitative results of frontal face synthesis are presented under various poses and occlusions. Then, the performance of face recognition on synthesized frontal-view and deoccluded faces is evaluated. Two categories of occlusions, including regular occlusion (i.e., square occluded block) and irregular occlusion, are tested. The experimental results on regular occlusion are provided and discussed in this section.

A. Experimental Settings

1) Databases: Multi-PIE [72] is the largest and widely used database for face recognition and synthesis experiments in constrained setting. A total of 750 000+ images from 337 subjects were recorded in four sessions; 20 illumination levels and 13 poses ranging from −90° to 90° are contained per subject. Following the testing protocol in [23] and [24], 337 subjects with neutral expression and 11 kinds of poses within ±75° from all sessions are used. The first 200 subjects formulate the training set, and the remaining 137 subjects are selected as the testing set. In the testing stage, the first image per subject with neutral illumination and frontal view is viewed as a gallery, and the remaining ones are considered as probes.

The LFW [73] database consists of 13 233 images of 5749 subjects, in which only 85 subjects have more than 15 images per subject and 4069 subjects have only one image per subject. The data set is usually used for testing face verification or synthesis performance in the wild (i.e., unconstrained setting). We follow the protocol of face verification [73], and the tenfold cross-validation strategy is exploited to evaluate the verification performance on the generated images. In addition, several SOTA models, such as FF-GAN [62], DR-GAN [23], and TP-GAN [24], are selected for comparison with our BoostGAN.

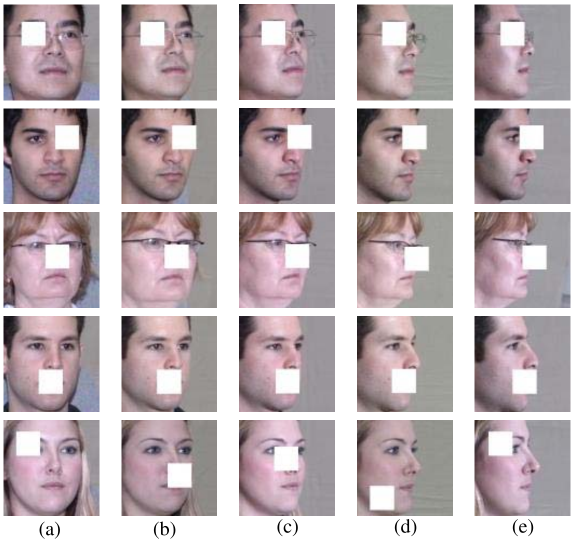

Fig. 4. Examples on regular block occluded Multi-PIE. From left to right, the poses are (a) 15°, (b) 30°, (c) 45°, (d) 60°, and (e) 75°. From top to bottom, the occluded masks are located on left eye, right eye, the tip of nose, the center of mouth, and random block, respectively.

2) Data Preprocessing: In data processing, both two data sets (i.e., Multi-PIE and LFW) are detected by using MTCNN [74] and aligned to a canonical view of size 128×128 so as to guarantee the generality of BoostGAN and reduce the model bias. Two kinds of regular occlusions, including keypoint position occlusion and random position occlusion in the square block, are considered in this article. For keypoint occlusion, the occlusion mask center is the facial key point, e.g., the left eye. For random occlusion, the occlusion mask center is randomly positioned. Each occlusion mask size is set as 32 × 32, filled with white pixels, such that the keypoint region can be completely covered. Fig. 4 shows some examples across different poses under regular occlusions, such as keypoint occlusion and random occlusion. Note that only the samples with key point occlusions from Multi-PIE are used as the training data for model training, which is then used for evaluating other data sets and occlusions.

3) Implementation Details: In conventional GAN, Goodfellow *et al.* [20] suggested to alternate between the k (usually $k = 1$) steps of optimizing D and the one step of optimizing G. Thus, we update the two steps for optimizing G^s and G^f, and 1 step for optimizing D, for performance guarantee. In all experiments, we set $\lambda_1 = 2e1$, $\lambda_2 = 4e1$, $\lambda_3 = 1$, $\lambda_4 = 3e-1$, and $\lambda_5 = 1e-3$. We train our network using the Adam optimizer [75]. The learning rate is fixed at 0.0001. The batch size is 4 and the training is stopped after $4 \sim 5$ epochs. Batch normalization and leaky ReLU are used in our model for accelerating convergence and avoiding model overfitting and gradient vanishing. Our model is trained on a NVIDIA 1080Ti GPU and implemented with TensorFlow. The whole process for model training takes about $6 \sim 8$ days.

In this section, the qualitative and quantitative experiments for image synthesis and recognition are divided into two parts: Multi-PIE under regular occlusion and LFW under regular occlusion. Each part includes the experimental results on keypoint and random occlusions under various poses. Two testing strategies, i.e., Ours and Ours* as shown in Fig. 3 with regard to different treatment of inputs, are employed in our experiments in order to bridge the real-world applications.

B. Face Frontalization and Recognition on Multi-PIE With Regular Occlusions

To qualitatively and quantitatively demonstrate the capability of our method for synthesis, the synthesized frontal images and recognition results under multiple kinds of occlusions and poses on Multi-PIE are presented in this section.

1) Face Frontalization and Recognition on Keypoint Occluded Multi-PIE: In order to demonstrate the impact of occlusion, the 32×32 white block occlusion is masked on each probe face. The block of this size is used to occlude one keypoint region (e.g., nose, eye, and mouth) in the 128×128 facial image. The synthetic results by using the existing DR-GAN [23] and TP-GAN [23] models trained on the nonoccluded Multi-PIE are shown in Fig. 1. We clearly observe that the synthetic quality of DR-GAN and TP-GAN* is seriously deteriorated when an occlusion exists. Note that, for TP-GAN, due to the missing pixels of the keypoint occlusion region, the global parametric model is employed. For clarity, we name this method as TP-GAN*.

The synthetic results on the keypoint position region occluded Multi-PIE are presented in Fig. 5, from which we can observe that the proposed BoostGAN model (i.e., Ours) can generate not only photorealistic but also identity-preserved faces under occlusions. This is significantly better than DR-GAN and TP-GAN. Especially, due to the lack of ground-truth frontal face supervision in DR-GAN, the hole (block occlusion) is still kept in the facial image without pixel filling in the hole. This excellent synthetic capability of our model benefits from the coarse-to-fine aggregation architecture, which can extract the complementary information from different inputs. Besides, the synthetic results with only one input of BoostGAN (i.e., Ours*) are also shown in Fig. 5. The performance is still better than DR-GAN and TP-GAN*.

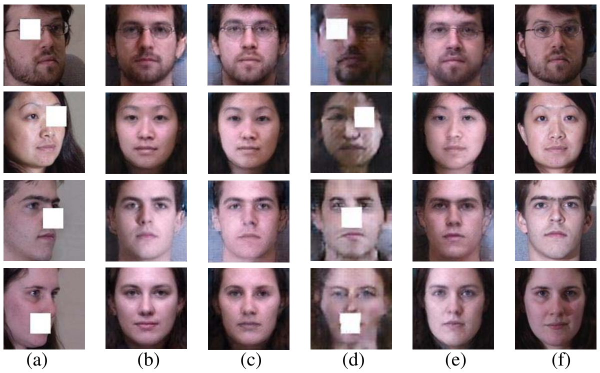

Fig. 5. Synthesis results on keypoint region occluded Multi-PIE data set. From top to bottom, the poses are 15°, 30°, 45°, 60°. The ground-truth frontal faces are provided at column (f). (a) Profile. (b) Ours. (c) Ours*. (d) [23]. (e) [24]*. (f) GT.

TABLE III

RANK-1 RECOGNITION RATE (%) COMPARISON ON KEYPOINT REGION OCCLUDED MULTI-PIE. BLACK: RANKS THE FIRST; RED: RANKS THE SECOND; BLUE: RANKS THE THIRD

Method	±15°	±30°	±45°	±60°	±75°
DR-GAN [23] (k1)	67.38	60.68	55.83	47.25	39.34
DR-GAN [23] (k2)	73.24	65.37	59.90	51.18	42.24
DR-GAN [23] (k3)	66.93	60.60	56.54	49.70	39.77
DR-GAN [23] (k4)	71.33	63.72	57.59	50.10	**40.87**
DR-GAN [23] (mean)	69.72	62.59	57.47	49.56	40.55
DR-GAN [23] (k3_DP)	54.9	54.9	53.3	50.7	-
TP-GAN [24]* (k1)	98.17	95.46	86.60	**65.91**	39.51
TP-GAN [24]* (k2)	**99.27**	**97.25**	**88.37**	**66.03**	**40.82**
TP-GAN [24]* (k3)	95.04	90.95	82.72	62.40	38.67
TP-GAN [24]* (k4)	97.80	93.66	83.84	62.27	36.76
TP-GAN [24]* (mean)	97.57	94.33	85.38	64.15	38.94
TP-GAN [24]* (k3_DP)	63.44	54.93	51.82	44.69	25.72
Ours*(k1)	99.03	96.21	**86.66**	64.45	39.74
Ours*(k2)	**99.12**	**96.33**	85.41	63.49	39.63
Ours*(k3)	96.03	92.06	83.10	63.45	39.92
Ours*(k4)	98.13	94.88	84.24	62.17	39.45
Ours*(mean)	98.08	94.87	84.85	63.39	39.69
Ours	**99.48**	**97.75**	**91.55**	**72.76**	**48.43**

The proposed BoostGAN model aims to synthesize and recognize faces under occlusions and pose variations. Therefore, for verifying the capability of different models for identity preservation, face recognition is studied on benchmark data sets. We first apply the trained GAN model to frontalize the profile faces, then extract the features of those generated frontal faces using the Light CNN, and finally evaluate the performance of face recognition or verification. Table III presents the recognition accuracies of different approaches, where ki, $i = 1, 2, 3, 4$ means the block mask region, such as left eye (k1), right eye (k2), nose (k3), and mouth (k4). We have the following observations:

1) It is obvious that by training the DR-GAN and TP-GAN with the keypoint occluded samples, i.e., DR-GAN(ki) and TP-GAN*(ki), the recognition performances are better than DR-GAN(k3_DP) and TP-GAN*(k3_DP) trained with deoccluded images. Note that DR-GAN (k3_DP) and TP-GAN*(k3_DP) denote that the model is trained on profile images without occlusions and tested on 32×32 *nose region* occluded Multi-PIE.

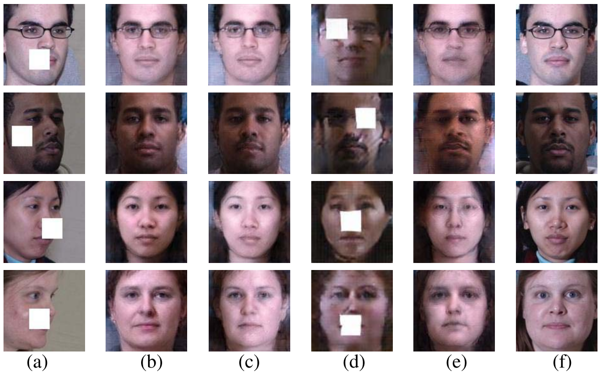

Fig. 6. Frontalization results on random block occluded Multi-PIE. The poses are 15°, 30°, 45°, and 60° from top to bottom. The GT frontal faces are provided at (f). Notably, all the models are trained solely on keypoint position occluded Multi-PIE data set. (a) Profile. (b) Ours. (c) Ours*. (d) [23]. (e) [24]*. (f) GT.

TABLE IV

RANK-1 RECOGNITION RATE (%) COMPARISON ON RANDOM BLOCK OCCLUDED MULTI-PIE. BLACK: RANKS THE FIRST; RED: RANKS THE SECOND; BLUE: RANKS THE THIRD

Method	±15°	±30°	±45°	±60°	±75°
DR-GAN [23] (r1)	47.64	38.93	33.21	25.38	18.92
DR-GAN [23] (r2)	65.75	55.15	46.52	38.33	29.00
DR-GAN [23] (r3)	56.01	46.27	39.13	29.11	23.01
DR-GAN [23] (r4)	59.10	47.92	39.97	33.69	25.20
DR-GAN [23] (mean)	57.13	47.07	39.71	31.63	24.03
TP-GAN [24]* (r1)	89.81	83.88	74.94	54.83	31.34
TP-GAN [24]* (r2)	77.98	71.68	60.52	42.68	23.92
TP-GAN [24]* (r3)	79.12	72.45	60.00	41.37	24.11
TP-GAN [24]* (r4)	86.13	77.76	64.84	45.08	25.15
TP-GAN [24]* (mean)	83.26	76.44	65.08	45.99	26.13
Ours*(r1)	**98.16**	**95.07**	86.67	**66.47**	**43.06**
Ours*(r2)	**98.10**	94.97	86.51	**66.56**	**42.96**
Ours*(r3)	98.02	94.99	**86.78**	66.25	42.69
Ours*(r4)	98.12	**95.12**	**86.69**	65.78	42.67
Ours*(mean)	**98.10**	95.04	86.66	66.27	42.85
Ours	**99.45**	**97.50**	**91.11**	**72.12**	**48.53**

2) The proposed BoostGAN outperforms DR-GAN and TP-GAN* for each type of block occlusion. It is common that with the increase of pose angles, the recognition accuracy is decreased. Also, with four same inputs in the testing of BoostGAN (i.e., Ours*) still outperforms DR-GAN and TP-GAN*, without performance degradation. Notably, with the same training set as ours, DR-GAN and TP-GAN* are all trained on Multi-PIE with keypoint occlusions, and the fairness in comparison is guaranteed.

3) The impact of occlusion is serious. This demonstrates three aspects. First, occlusion is an important impact which deteriorates face synthesis and recognition. Second, the existing GAN variants cannot well deal with occlusions. Third, GAN models for face frontalization in real-world scenarios are worthy to be noticed.

2) Face Frontalization and Recognition on Random Block Occluded Multi-PIE: Based on the trained model on the data set with keypoint position region occlusions, the test results on profile images with random occlusions are shown in Fig. 6. Obviously, the synthesized faces of BoostGAN with two different testing schemes are still significantly better than both DR-GAN and TP-GAN*. The generated faces of TP-GAN* are distorted due to that the randomly occluded profile images may not appear in the training stage. In addition, due to the randomness of occlusion, the position of occlusion in the generated images by DR-GAN is also changed. However, different from them, the proposed method can also obtain good synthetic quality. Furthermore, by increasing the pose angle, BoostGAN can enable the frontal-view face synthesis with clearer and cleaner details.

The rank-1 recognition rates for the random block occlusions are shown in Table IV, where r$i, i = 1, 2, 3, 4$ denotes the random block occlusion mask. It can be seen that the recognition performance of our two models with different testing schemes is significantly better than others. Especially, by comparison between Tables III and IV, we observe a dramatically decreased recognition rates of DR-GAN and TP-GAN* due to the changes of occlusions from keypoint to random type. However, the proposed BoostGAN still keeps stable good performance and the robustness to different occlusion type is verified.

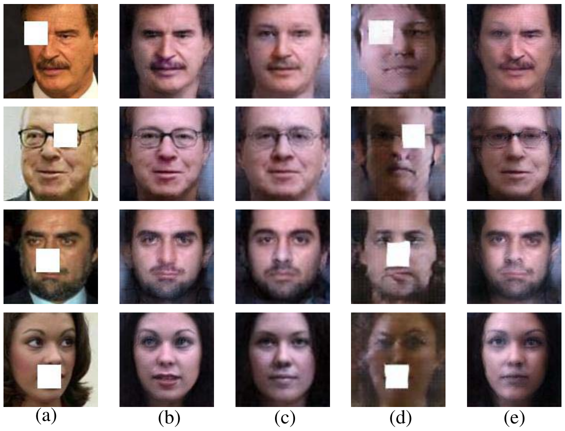

Fig. 7. Frontalization results on keypoint region occluded LFW data set in the wild. Notably the ground-truth frontal images for this data set are unavailable. The models are trained based solely on keypoint position occluded Multi-PIE data set. (a) Profile. (b) Ours. (c) Ours*. (d) [23]. (e) [24]*.

C. Face Frontalization and Verification on LFW With Regular Occlusions

1) Face Frontalization and Verification on Keypoint Occluded LFW: In order to demonstrate the generalization ability of the model in the wild, the LFW database is used to test the model. As shown in Fig. 7, our BoostGANs also obtain superior visual performance on LFW to others, but the synthesized background color is similar to that of Multi-PIE. This is understandable and generally happens for all models because the models are only trained on Multi-PIE.

The face verification performance is evaluated by the recognition accuracy (ACC) and AUC, and the results are reported in Table V. From the results, we see that DR-GAN fails due to its weak specificity to occlusions. As expected, both our models still keep the superior performance compared with DR-GAN and TP-GAN*.

TABLE V

FACE VERIFICATION ACCURACY (ACC) AND AUC RESULTS ON KEYPOINT REGION OCCLUDED LFW

Method	ACC(%)	AUC(%)
DR-GAN [23] (k1)	67.60	73.65
DR-GAN [23] (k2)	67.28	72.94
DR-GAN [23] (k3)	58.43	59.19
DR-GAN [23] (k4)	69.50	76.05
DR-GAN [23] (mean)	65.71	70.46
TP-GAN [24]* (k1)	86.52	92.81
TP-GAN [24]* (k2)	87.83	93.96
TP-GAN [24]* (k3)	85.17	91.63
TP-GAN [24]* (k4)	87.78	**93.97**
TP-GAN [24]* (mean)	86.83	93.09
Ours*(k1)	**88.47**	**94.07**
Ours*(k2)	**88.13**	93.85
Ours*(k3)	86.62	93.66
Ours*(k4)	87.88	93.92
Ours*(mean)	87.78	93.88
Ours	**89.57**	**94.90**

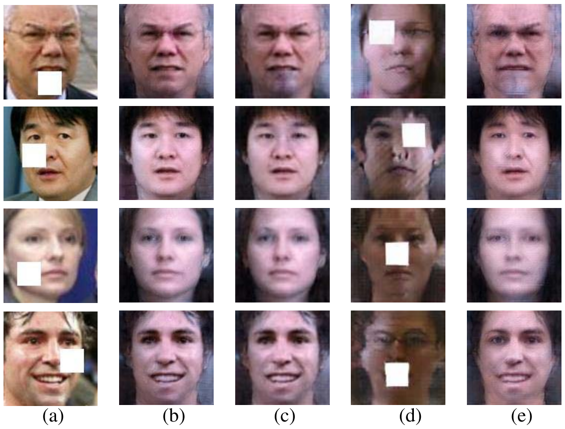

Fig. 8. Frontalization results on random block occluded LFW data set. Notably, all the models are trained solely on keypoint position occluded Multi-PIE data set, without retraining on randomly blocked data sets. (a) Profile. (b) Ours. (c) Ours*. (d) [23]. (e) [24]*.

2) Face Frontalization and Verification on Random Block Occluded LFW: The qualitative and quantitative experimental results are also tested on random block occluded LFW, as is shown in Fig. 8 and Table VI, respectively. We can observe that the synthetic quality of frontal-view facial images of our BoostGANs is better than others.

From Table VI, we can see that the verification performance of our models is much superior to others. Furthermore, TP-GAN has shown comparable results as presented in Table V under keypoint position occlusion but significantly degraded accuracy as presented in Table VI under random position occlusion in the wild. Similar to constrained Multi-PIE data set, our proposed models show an SOTA performance under random occlusions in the unconstrained scenario, and there is also no performance degradation across different types of occlusion, i.e., from keypoint type to random type. Hence, we could conclude that our proposed models show good generalization ability and robust performance under regular occlusions in constrained and unconstrained scenarios.

TABLE VI

FACE VERIFICATION ACCURACY (ACC) AND AUC RESULTS ON RANDOM BLOCK OCCLUDED LFW

Method	ACC(%)	AUC(%)
DR-GAN [23] (r1)	63.28	67.20
DR-GAN [23] (r2)	65.53	71.79
DR-GAN [23] (r3)	57.15	57.76
DR-GAN [23] (r4)	64.82	70.35
DR-GAN [23] (mean)	62.70	66.78
TP-GAN [24]* (r1)	82.75	89.86
TP-GAN [24]* (r2)	77.65	84.63
TP-GAN [24]* (r3)	81.07	88.24
TP-GAN [24]* (r4)	83.25	90.15
TP-GAN [24]* (mean)	81.18	88.22
Ours*(r1)	**87.92**	**93.46**
Ours*(r2)	86.78	92.81
Ours*(r3)	87.28	93.15
Ours*(r4)	87.18	**93.26**
Ours*(mean)	**87.29**	93.17
Ours	**89.58**	**94.75**

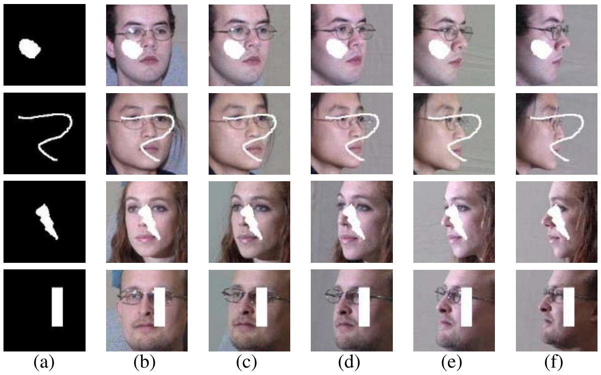

Fig. 9. Examples on irregular block occluded Multi-PIE. (a) Four kinds of irregular masks deployed in this article. (b)–(f) Irregular block occluded face samples under the poses of ±15°, ±30°, ±45°, ±60°, and ±75° based on the four masks. (a) Mask. (b) ±15°. (c) ±30°. (d) ±45°. (e) ±60°. (f) ±75°.

For further verifying the performance of our model under irregular occlusions, in the following, extensive experiments for testing on irregular occlusions are deployed.

V. EXPERIMENTS UNDER IRREGULAR OCCLUSIONS

In real-world scenarios, the occlusions on the facial image are often irregular. Therefore, for testing on a more realistic scenario, we consider several types of irregular occlusions. In the following, we provide the test results of synthetic and recognition/verification results on Multi-PIE and LFW with irregular occlusions. The trained model is still based on the data with regular occlusions.

A. Testing Data Processing

Four different types of irregular mask M are presented for occlusion, as shown in Fig. 9(a). The size of the irregular mask M is 128 × 128, which is the same as the input image. The mask of M is encoded with binary value, i.e., $M(i, j) \in \{0, 1\}$, where (i, j) denotes the coordinate of pixel. $M(i, j) = 0$ indicates the pixel at (i, j) of a facial image is lost, while $M(i, j) = 1$ indicates the pixel at (i, j) is preserved.

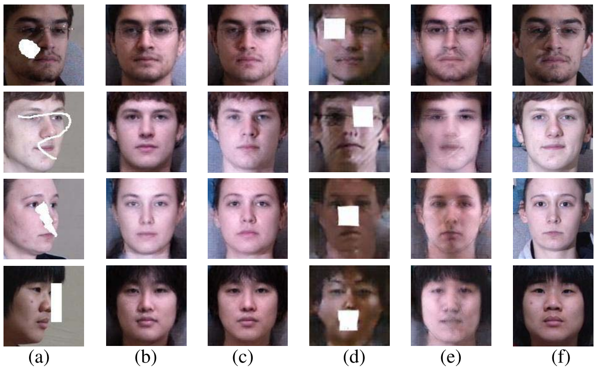

Fig. 10. Frontalization results on irregular block occluded Multi-PIE. The poses are 15°, 30°, 45°, and 60° from top to bottom. The GT frontal images are provided at (f). Notably, all the models are trained solely on keypoint position occluded Multi-PIE data set. (a) Profile. (b) Ours. (c) Ours*. (d) [23]. (e) [24]*. (f) GT.

TABLE VII
COMPARISON OF RANK-1 RECOGNITION RATE (%) ON IRREGULAR BLOCK OCCLUDED MULTI-PIE. BLACK: RANKS THE FIRST; RED: RANKS THE SECOND; BLUE: RANKS THE THIRD

Method	±15°	±30°	±45°	±60°	±75°
DR-GAN [23] (m1)	60.51	52.68	44.88	35.69	23.59
DR-GAN [23] (m2)	73.40	63.59	50.74	40.10	29.13
DR-GAN [23] (m3)	73.38	65.44	55.63	44.16	33.56
DR-GAN [23] (m4)	63.70	43.62	35.74	32.17	27.34
DR-GAN [23] (mean)	67.75	56.33	46.75	38.03	28.41
TP-GAN [24]* (m1)	94.11	89.98	81.32	57.23	28.93
TP-GAN [24]* (m2)	61.60	48.65	39.19	24.47	13.98
TP-GAN [24]* (m3)	86.21	79.40	67.69	47.69	24.34
TP-GAN [24]* (m4)	85.84	76.71	64.97	44.35	26.18
TP-GAN [24]* (mean)	81.94	73.69	63.29	43.44	23.36
Ours*(m1)	**99.12**	**96.84**	**90.20**	**70.36**	**44.29**
Ours*(m2)	93.27	86.13	70.92	46.08	27.65
Ours*(m3)	93.14	88.17	76.53	56.67	34.52
Ours*(m4)	**98.56**	**95.21**	**85.99**	**65.59**	**42.11**
Ours*(mean)	96.02	91.59	80.91	59.68	37.14
Ours	**98.81**	**96.10**	**88.14**	**67.23**	**44.03**

For testing examples with occlusions, by multiplying these four irregular masks (i.e., m1–m4) with facial image, we can obtain four groups of data with different irregular occlusions in experiments, which is shown in Fig. 9(b)–(f).

B. Face Frontalization and Recognition on Multi-PIE With Irregular Occlusions

The qualitative synthetic results with frontal-view faces generated by different models on Multi-PIE with irregular occlusions are shown in Fig. 10. It can be seen that the proposed BoostGANs show better realistic image quality than others. Other models show obvious artifacts and ghosting effects. Specifically, when only one irregular occlusive profile is prepared, the synthetic quality of our proposed method BoostGAN* is still much better than DR-GAN and TP-GAN*. Furthermore, if multiple images with irregular occlusions are prepared, better realistic quality can be achieved by BoostGAN, as shown in the second column of Fig. 10. It illustrates that the generalization ability and robustness of our approach are more excellent than other methods.

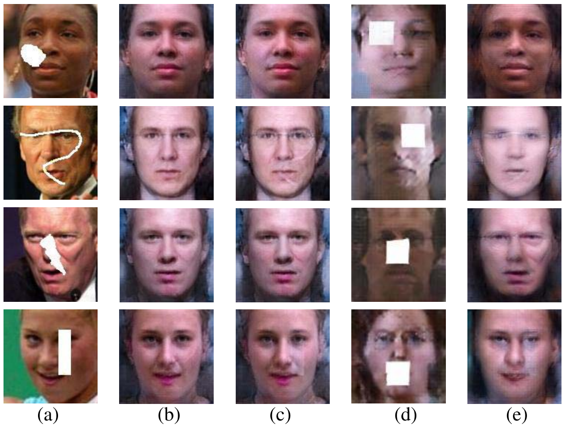

Fig. 11. Frontalization results on irregular block occluded LFW data set. Notably, all the models are trained solely on keypoint position occluded Multi-PIE data set, without retraining on irregular occluded data sets. (a) Profile. (b) Ours. (c) Ours*. (d) [23]. (e) [24]*.

We further quantitatively compare our proposed approach with other models in terms of recognition accuracy on irregular occluded Multi-PIE, which is shown in Table VII. Table VII shows the rank-1 accuracies of different methods under irregular occlusions in the constrained Multi-PIE data set. The recognition accuracies of our proposed BoostGANs significantly outperform the state-of-the-art TP-GAN and DR-GAN models across all poses. The robustness of BoostGAN to occlusions with different shapes is clearly demonstrated.

C. Face Frontalization and Verification on LFW With Irregular Occlusions

In the unconstrained LFW data set, the qualitative synthetic images are shown in Fig. 11. From the visual quality of frontal-view faces, the performance of the proposed BoostGANs still surpass other models and our model can generate more photorealistic frontal-view faces with better texture details, while other models show obvious distortion and color effect. The photorealism of BoostGAN is much improved.

The quantitative experiments on the synthetic frontal-view faces are also conducted, and the ACC and AUC results are reported in Table VIII. From the results, we can observe that our proposed method shows much better performance than other methods. The proposed BoostGAN is insusceptible to different occlusions in constrained and unconstrained scenarios, and the advantages over other SOTA GAN models are empirically demonstrated.

It is noteworthy that all the experiments with irregular occlusions are tested on the well-trained model with the Multi-PIE database under keypoint occlusions. The excellent performance of BoostGAN under irregular occlusion in a controlled and uncontrolled setting fully reflects the outstanding generalization ability and robustness of our proposed approach. Therefore, the proposed BoostGAN has shown a more practical application value in face recognition under unrestricted environments, when pose variations and occlusions co-exist. The mind of boosting model in this article may also be adapted to other application scenarios, e.g., image-to-image translation, synthesis of expression, etc.

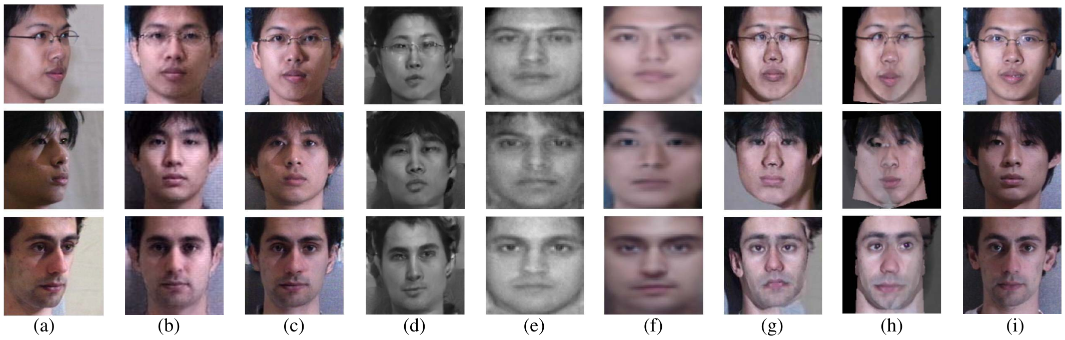

Fig. 12. Comparison with SOTA synthesis models under the pose variation of 45° (the first two rows) and 30° (the third row). Our BoostGAN model is trained from scratch on the nonoccluded Multi-PIE data set. (a) Profile. (b) Ours. (c) [24]. (d) [23]. (e) [76]. (f) [77]. (g) [60]. (h) [21]. (i) GT.

TABLE VIII

FACE VERIFICATION ACCURACY (ACC) AND AUC RESULTS IRREGULAR BLOCK OCCLUDED LFW

Method	ACC(%)	AUC(%)
DR-GAN [23] (m1)	66.27	71.85
DR-GAN [23] (m2)	66.93	73.11
DR-GAN [23] (m3)	58.22	59.20
DR-GAN [23] (m4)	66.82	72.58
DR-GAN [23] (mean)	64.56	69.19
TP-GAN [24]* (m1)	84.17	91.22
TP-GAN [24]* (m2)	77.12	83.88
TP-GAN [24]* (m3)	82.80	89.87
TP-GAN [24]* (m4)	84.32	91.03
TP-GAN [24]* (mean)	82.10	89.00
Ours*(m1)	**88.45**	**94.09**
Ours*(m2)	84.98	91.59
Ours*(m3)	86.23	92.83
Ours*(m4)	**86.98**	**93.06**
Ours*(mean)	86.66	92.89
Ours	**88.52**	**93.91**

VI. EXPERIMENTS OF TWO-STEP METHOD WITH OCCLUSION REMOVAL

A. Motivation

It is natural to think of that the problem of profile face frontalization and recognition with occlusions can be solved by a general two-step method: 1) deocclusion and 2) face frontalization. For the first step, facial deocclusion can be done through image completion methods. Then, for the second step, face frontalization approaches (e.g., TP-GAN) can be used to rotate the clean profile facial image to frontal facial image. However, several problems will be encountered for image completion methods, which are given in the following.

1) The image after deocclusion by using the existing image completion methods will lose texture details and identity information, which cannot be improved by the subsequent face frontalization approaches.
2) As mentioned in Section I, the previous image completion methods are more feasible to close-set image completion task, and the image inpainting aims to restore the really fine details and semantic facial structures, rather than to exactly regress the ground truth.
3) Those face completion methods aim to restore the frontal or near-frontal images instead of profile images. With these considerations, the two-step method may not be suitable for solving the problem of face frontalization and recognition under occlusions.

TABLE IX

RANK-1 RECOGNITION RATE (%) COMPARISON ON PROFILE MULTI-PIE WITH OCCLUSION REMOVED FIRST

Method	±15°	±30°	±45°	±60°	mean
FIP+LDA [78]	90.7	80.7	64.1	45.9	70.35
MVP+LDA [79]	92.8	83.7	72.9	60.1	77.38
CPF [76]	95.0	88.5	79.9	61.9	81.33
DR-GAN [23]	94.0	90.1	86.2	83.2	88.38
DR-GAN$_{AM}$ [29]	95.0	91.3	88.0	**85.8**	90.03
FF-GAN [62]	94.6	92.5	89.7	85.2	**90.50**
TP-GAN [24]	**98.68**	**98.06**	**95.38**	**87.72**	**94.96**
Light CNN [70]	**98.59**	**97.38**	**92.13**	62.09	87.55
Ours	**99.88**	**99.19**	**96.84**	**87.52**	**95.86**

B. Experimental Results

In order to avoid the abovementioned issues and demonstrate the superiority of our proposed method without occlusions to others, we consider a perfect deocclusion, that is, we intuitively adopt the clean profile images without occlusions for experiments. Face synthesis and recognition results on the Multi-PIE database are provided and shown in Fig. 12 and Table IX, respectively. We have the following observations.

1) *Face Synthesis on Multi-PIE After De-Occlusion:* The synthetic results of BoostGAN are compared against SOTA face frontalization models on clean profile faces after deocclusion, that is, the deoccluded Multi-PIE represents the original but clean Multi-PIE. Fig. 12 shows the generated faces, from which we can see that the generated results by BoostGAN still show the competitive performance and superiority to TP-GAN but obviously outperforms other approaches no matter in global structure or local texture. This validates that BoostGAN can also work well under nonoccluded scenario where the occlusion is supposed to be removed.
2) *Face Recognition on Multi-PIE After Deocclusion:* We further verify the effectiveness of the proposed method in experiment on clean profile faces. Specifically, the rank-1 accuracies of different models are presented in Table IX. Specifically, eight methods, including FIP+LDA [78], MVP+LDA [79], CPF [76], DR-GAN [23], DR-GAN$_{AM}$ [29], FF-GAN [62], and TP-GAN [24], are fairly compared. The results of Light CNN are denoted as baselines. For a fair comparison, all the methods follow the same experimental settings.

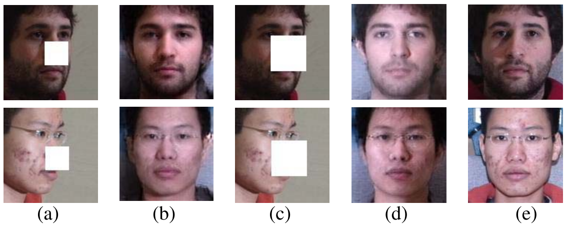

Fig. 13. Synthesis results on different size of nose block occlusion. (a) and (c) From top to bottom, the poses are 30° and 60°, respectively. (b) Synthesis result of (a) with 32 × 32 occlusion. (d) Synthesis result of (c) with 48 × 48 occlusion. (e) GT frontal faces. Notably, all the models are trained solely on keypoint position occluded Multi-PIE data set. Ours* means that four same images with occlusion are fed as the input (a) Profile. (b) Ours*. (c) Profile. (d) Ours*. (e) GT.

TABLE X

RANK-1 RECOGNITION RATE (%) WITH DIFFERENT SIZE OF NOSE-OCCLUSION. BLACK: RANKS THE FIRST AND RED: RANKS THE SECOND

Method	±15°	±30°	±45°	±60°	±75°
Ours*(k3)	**96.03**	**92.06**	**83.10**	**63.45**	**39.92**
Ours*(k3_48)	**77.32**	**74.52**	**65.59**	**46.36**	**26.98**

We clearly observe that the proposed BoostGAN is superior to all other methods in recognizing clean profile faces. TP-GAN [24], as the SOTA model in clean face frontalization, is also inferior to ours. The average accuracy across multiple poses of BoostGAN is 95.86%, which is higher than TP-GAN with 94.96% even though the occlusion does not exist.

The experiments demonstrate that even though the occlusions are removed, the synthesis and recognition performances of our method still surpass others. Note that, in order to have an insight into the training effectiveness of irregular occlusions, we have further conducted another study by adding irregular occlusions into regular occlusions for model training. The results and discussion are placed in Section VII-E.

VII. MODEL ANALYSIS AND DISCUSSION

A. Analysis of Different Size of Occlusion Mask

This section discusses the impact of the size of occlusion mask to face frontalization and recognition performance. By frozen the well-trained BoostGAN model on the training data under 32 × 32 occlusion, the model is tested on the faces with 32 × 32 and 48 × 48 keypoint region occlusions (nose region). The qualitative and quantitative results are presented in Fig. 13 and Table X, respectively. From the results, we can see that the performance of BoostGAN is reasonably degraded with the increasing size of occlusion mask. This is understandable that differently sized occlusion shows different structural information and large occlusion can cause performance degradation. It is also known that if the whole face is occluded, the models fail to recognize what the faces show.

B. Analysis of Parameter Sensitivity

In this section, we analyze the effects of the five tradeoff parameters λ_1–λ_5 in model (10). The recognition accuracy with respect to each parameter by fixing others as default values is shown in Fig. 14. The sensitivity to the tradeoff parameters of our model is explicitly observed. We also see that with the change in λ_1, λ_4, and λ_5, the recognition accuracy changes slightly. It shows that the L_{adv}, L_{sym}, and L_{tv} losses have minor effects on the identity preservation of the generated image. However, when changing the values of λ_2 and λ_3 far away from the optimum values, the accuracy drops sharply. It means that the L_{ip} and L_{pix} losses play a key role in identity preservation of synthesized frontal facial image under occlusions. These results suggest that each loss function in BoostGAN is essential for photorealistic and identity preservation during synthesis.

TABLE XI

COMPARISON WITH DIFFERENT BOOSTING ACROSS COARSE AND FINE GENERATOR ON KEYPOINT REGION OCCLUDED MULTI-PIE FOR BETTER INSIGHT OF THE BOOSTING EFFECT

Boosting	±15°	±30°	±45°	±60°	±75°
Only 1 fine	95.52	86.20	69.29	46.99	32.18
1 coarse + 1 fine	98.47	93.13	78.16	53.93	34.70
2 coarse + 1 fine	98.95	95.53	86.09	61.62	36.97
3 coarse + 1 fine	99.39	96.23	85.20	60.67	39.91
Only 4 coarse	97.63	90.78	76.14	51.38	33.18
4 coarse + 1 fine	**99.48**	**97.75**	**91.55**	**72.76**	**48.43**

C. Ablation Analysis of Boosting Generator

In order to demonstrate the effectiveness of the boosting generator, the ablation analysis of boosting variants is provided in Table XI, in which "coarse" and "fine" denote the coarse image generated by the multiocclusion frontal view generator G^c and the fine image synthesized by the multi-input boosting network G^f, respectively. For example, "1 coarse + 1 fine" means that only one coarse image and one fine image are used to optimize the parameters of BoostGAN in the training stage. Note that "4 coarse + 1 fine" represents our proposed BoostGAN. Specifically, from the results listed in Table XI, we have the following observations.

1) With the increase of the number of intermediated coarse images, the recognition performance improves rapidly. It means that the boosting mechanism proposed in this article for an ensemble of the intermediated images generated by the coarse generator can improve the identity preserving ability of finally synthesized image.
2) In the case of "Only 1 fine," the intermediated images generated by the coarse generator (i.e., G^c) do not participate the optimization in the training process. In other words, G^c is not trained in this case which can be labeled as Ours (w/o G^c). We can see that the result of "4 coarse + 1 fine" is far better than those of "Only 1 fine," which demonstrates that the intermediated images generated by the coarse generator G^c can provide more complementary identity information and the proposed boosting generator (i.e., fine generator) is beneficial to better facial synthesis by ensemble of the intermediate coarse images.
3) In the case of "Only 4 coarse," the proposed boosting generator G^f is not used in the training phase, which can be labeled as Ours (w/o G^f), and one of the coarse

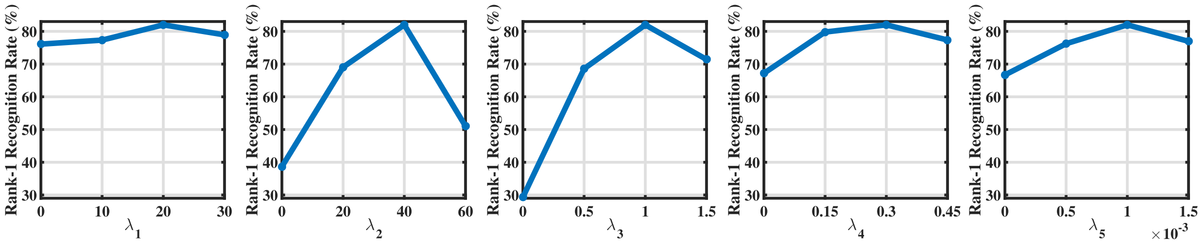

Fig. 14. Rank-1 recognition rate (%) on keypoint region occluded Multi-PIE with respect to λ_1–λ_5.

TABLE XII

COMPARISON WITH THE PERFORMANCE WITHOUT ADVERSARIAL LEARNING IN THE INTERMEDIATE GENERATED IMAGES ON THE KEYPOINT REGION OCCLUDED MULTI-PIE

Method	±15°	±30°	±45°	±60°	±75°
Ours (w/o IAL)	99.03	96.44	87.85	64.10	38.36
Ours	**99.48**	**97.75**	**91.55**	**72.76**	**48.43**

TABLE XIII

COMPARISON OF RANK-1 RECOGNITION RATE (%) ON IRREGULAR BLOCK OCCLUDED MULTI-PIE WITH DIFFERENT KINDS OF TRAINING OCCLUSIONS

Training Occlusion	±15°	±30°	±45°	±60°	±75°
4 Regular	98.81	96.10	88.14	67.23	44.03
3 Reg. + 1 Irreg.	**99.59**	97.87	91.98	70.58	45.69
4 Irregular	99.58	**98.14**	**93.95**	**77.61**	**52.91**

images from G^c is tested. Obviously, the recognition performance of "4 coarse + 1 fine" is still far better than "Only 4 coarse." The results clearly demonstrate the effectiveness of the boosting generator for aggregation over each of the four coarse images.

4) By comparing the results of "Only 1 fine" with "Only 4 coarse," we can see the recognition performance of "Only 4 coarse" is better than "Only 1 fine." This is easy to understand that the coarse generator is a deeper network than the boosting fine generator, and the coarse generation is important for identity preservation. The boosting generator is beneficial to the deocclusion and photorealistic image generation.

The abovementioned analysis fully demonstrates the effectiveness of the boosting generator in the BoostGAN model.

D. Analysis of Intermediate Adversarial Learning

For illustrating the effectiveness of intermediate adversarial learning (IAL), the recognition performance of BoostGAN without IAL which is denoted as Ours (w/o IAL) is provided, as shown in Table XII. We clearly observe that the recognition result of BoostGAN (i.e., Ours) is better than that of "Ours (w/o IAL)." This is because the excellent synthetic performance of GAN relies on the game between generator and discriminator. Without the IAL, the synthetic quality of G^c may become worse. As presented in Table XI, we know that the coarse generated images have an impact on the final performance. Therefore, without the IAL, the coarse images may be destroyed, which accordingly affects the synthetic quality of the final results generated by G^f. It can be concluded that the IAL is undoubtedly important for improving the performance.

E. Training With Both Regular and Irregular Occlusions

In the generation experiments under irregular occlusions presented in Section V, the synthesis results are obtained by the models trained solely on the keypoint position occluded Multi-PIE data set (i.e., "Regular" occlusion), without including or fine-tuning on the irregular occlusions. In this section, we try to add the irregular occluded Multi-PIE into the training data together with regular occlusions. The recognition results trained with different kinds of occlusions of images in the training set are provided in Table XIII. For example, "3 Reg. + 1 Irreg." means that three keypoint region occluded profile images (regular block) and one irregular block occluded profile image are fed as the input of BoostGAN in the training phase. Note that the test data are composed of the irregular block occluded images. Obviously, the recognition performances of "3 Regular + 1 Irregular" and "4 Irregular" are better than "4 Regular." Particularly, by only using the irregular occlusions for training, the performance is far better than others on the large-pose profile images. It is easy to understand that by adding the irregular occlusions for training, the distribution gap between the training and testing data is reduced such that the performance on irregular occlusion is further improved.

Besides, we further test the abovementioned three trained models under different kinds of occlusions on the keypoint region (regular block) occluded profile images, and the recognition results are presented in Table XIV. It can be seen that the recognition performance of "4 Regular" is much better than that of "3 Regular + 1 Irregular" and "4 Irregular." The reason is the same as that given in Table XIII, that is, the distribution discrepancy between training and testing is much smaller. It is worth noting that although the training set is changed, the recognition performances do not drop sharply. It demonstrates the robustness of our proposed method.

F. Discussion of the Advantage of BoostGAN

Our approach is recognized to be a boosting model to achieve automatic deocclusion and frontalization of faces simultaneously, which aims at recognizing faces when occlusions and large-pose variations coexist. Although BoostGAN succeeds in the understudied face frontalization under occlusion, some similar modules with the existing GAN variants are deployed. First, the CNN-based encoder–decoder architecture is used. Second, the pixel- and feature-level loss functions are exploited. Third, the basic parts, such as G and D in GAN, are considered. The main difference of our model from other GAN variants is that the complementary information from

TABLE XIV

RANK-1 RECOGNITION RATE (%) COMPARISON ON KEYPOINT REGION (REGULAR BLOCK) OCCLUDED MULTI-PIE WITH DIFFERENT KINDS OF TRAINING OCCLUSIONS

Training Occlusion	±15°	±30°	±45°	±60°	±75°
4 Regular	**99.48**	**97.75**	**91.55**	**72.76**	**48.43**
3 Reg. + 1 Irreg.	99.00	96.17	89.20	66.03	38.55
4 Irregular	98.63	94.43	81.64	53.99	35.26

multiple inputs is boosted through a coarse-to-fine generator. The structure of BoostGAN is simple, intuitive but effective for occlusion- and pose-aware face frontalization and recognition.

VIII. CONCLUSION

This article contributes to the answer of how to recognize faces if occlusion and large-pose variation coexist simultaneously. In the first place, we raise this question and prove the deficiency of existing methods. To look more into the occlusion, in this article, we contribute a BoostGAN model for the pose- and occlusion-aware frontalization and recognition of faces in constrained and unconstrained scenarios. The proposed method is an end-to-end framework equipped with a coarse-to-fine face deocclusion and frontalization network ensemble. The coarse generator is used to achieve coarse frontalization and deocclusion across multiple occlusions and large-pose variations. The boosting network aims to generate clean, frontal, and photorealistic faces with identity preservation by boosting the complementary information of multiple inputs. The generality and superiority of BoostGAN are validated by extensive experiments on benchmark data sets and outperforms other SOTA GAN models under different types of occlusions, including regular and irregular occlusions. In addition, we have provided further experimental analysis and discussion of our BoostGAN from the perspective of model design.

This article focuses on automatic deocclusion by leveraging the capability of synthesis of GAN models from multiple inputs with complementary information. The prior of occlusion is not explicitly considered; therefore, the future work by integrating prior modeling on the occlusion into GAN networks will be a challenge.

REFERENCES

[1] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. CVPR*, 2014, pp. 1891–1898.

[2] M. Shao, Y. Zhang, and Y. Fu, "Collaborative random faces-guided encoders for pose-invariant face representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 1019–1032, Apr. 2018.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2014, pp. 815–823.

[4] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. CVPR*, 2014, pp. 212–220.

[5] H. Wang *et al.*, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. CVPR*, Jun. 2018, pp. 5265–5274.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, 2005, pp. 886–893.

[7] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. ICCV*, 1999, pp. 1150–1157.

[8] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[9] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 2, no. 7, pp. 1160–1169, Jul. 1985.

[10] J. Yang, L. Zhang, Y. Xu, and J.-Y. Yang, "Beyond sparsity: The role of L1-optimizer in pattern classification," *Pattern Recognit.*, vol. 45, no. 3, pp. 1104–1118, Mar. 2012.

[11] J. Yang, D. Chu, L. Zhang, Y. Xu, and J. Yang, "Sparse representation classifier steered discriminative projection with applications to face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 7, pp. 1023–1035, Jul. 2013.

[12] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[13] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. NIPS*, 2014, pp. 1988–1996.

[14] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," in *Proc. ICCV*, 2013, pp. 1489–1496.

[15] M. Kan, S. Shan, and X. Chen, "Multi-view deep network for cross-view classification," in *Proc. CVPR*, 2016, pp. 4847–4855.

[16] K. Cao, R. Yu, L. Cheng, X. Tang, and C. L. Chen, "Pose-robust face recognition via deep residual equivariant mapping," in *Proc. CVPR*, 2018, pp. 5187–5196.

[17] Q. Duan, L. Zhang, and W. Zuo, "From face recognition to kinship verification: An adaptation approach," in *Proc. ICCVW*, 2017, pp. 1590–1598.

[18] Q. Duan and L. Zhang, "AdvNet: Adversarial contrastive residual net for 1 million kinship recognition," in *Proc. ACMMMW*, 2017, pp. 21–29.

[19] Q. Duan, L. Zhang, and W. Jia, "Adv-kin: An adversarial convolutional network for kinship verification," in *Proc. CCBR*, 2017, pp. 1–10.

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

[21] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. CVPR*, 2015, pp. 4295–4304.

[22] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *Proc. ICCV*, 2015, pp. 3871–3879.

[23] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. CVPR*, 2017, pp. 1415–1424.

[24] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proc. ICCV*, 2017, pp. 2439–2448.

[25] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proc. CVPR*, 2018, pp. 8398–8406.

[26] Z. Zhang, X. Chen, B. Wang, G. Hu, W. Zuo, and E. R. Hancock, "Face frontalization using an appearance-flow-based convolutional neural network," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2187–2199, May 2019.

[27] P. Li, X. Wu, Y. Hu, R. He, and Z. Sun, "M2FPA: A multi-yaw multi-pitch high-quality dataset and benchmark for facial pose analysis," in *Proc. ICCV*, 2019, pp. 10043–10051.

[28] C. Fu, Y. Hu, X. Wu, G. Wang, Q. Zhang, and R. He, "High fidelity face manipulation with extreme pose and expression," 2019, *arXiv:1903.12003*. [Online]. Available: http://arxiv.org/abs/1903.12003

[29] L. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 3007–3021, Dec. 2019.

[30] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," in *Proc. CVPR*, 2003, pp. 1–6.

[31] A. Levin, A. Zomet, and Y. Weiss, "Learning how to inpaint from global image statistics," in *Proc. ICCV*, 2003, p. 305.

[32] Y. Wexler, E. Shechtman, and M. Irani, "Space-time completion of video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 463–476, Mar. 2007.

[33] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. CVPR*, 2017, pp. 6721–6729.

[34] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. ECCV*, 2018, pp. 85–100.

[35] Y. Song *et al.*, "Contextual-based image inpainting: Infer, match, and translate," in *Proc. ECCV*, 2018, pp. 3–19.

[36] 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.

[37] 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.

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

[39] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Jul. 2017.

[40] Z. Chen, S. Nie, T. Wu, and C. G. Healey, "High resolution face completion with multiple controllable attributes via fully end-to-end progressive generative adversarial networks," 2018, *arXiv:1801.07632*. [Online]. Available: http://arxiv.org/abs/1801.07632

[41] Y. Zhao *et al.*, "Identity preserving face completion for large ocular region occlusion," in *Proc. BMVC*, 2018.

[42] Q. Duan and L. Zhang, "BoostGAN for occlusive profile face frontalization and recognition," 2019, *arXiv:1902.09782*. [Online]. Available: http://arxiv.org/abs/1902.09782

[43] A. Dosovitskiy, J. T. Springenberg, and T. Brox, "Learning to generate chairs with convolutional neural networks," in *Proc. CVPR*, 2015, pp. 1538–1546.

[44] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017, pp. 1125–1134.

[45] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Proc. ECCV*, 2016, pp. 318–333.

[46] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. CVPR*, 2017, pp. 4681–4690.

[47] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. ICLR*, 2019.

[48] L. Zhang, L. Lin, X. Wu, S. Ding, and L. Zhang, "End-to-end photo-sketch generation via fully convolutional representation learning," in *Proc. ICMR*, 2015, pp. 627–634.

[49] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. ICCV*, 2017, pp. 1501–1510.

[50] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "CartoonGAN: Generative adversarial networks for photo cartoonization," in *Proc. CVPR*, 2018, pp. 9465–9474.

[51] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. CVPR*, 2017, pp. 5810–5818.

[52] Y. Li, L. Song, X. Wu, R. He, and T. Tan, "Anti-makeup: Learning a bi-level adversarial network for makeup-invariant face verification," in *Proc. AAAI*, 2018.

[53] Y. Shen, P. Luo, J. Yan, X. Wang, and X. Tang, "FaceID-GAN: Learning a symmetry three-player GAN for identity-preserving face synthesis," in *Proc. CVPR*, 2018, pp. 821–830.

[54] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. ICLR*, 2016.

[55] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," in *Proc. ICML*, 2017.

[56] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*. [Online]. Available: http://arxiv.org/abs/1802.05957

[57] 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. 2223–2232.

[58] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. ICLR*, 2018.

[59] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng, "3D-aided dual-agent GANs for unconstrained face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2380–2394, Oct. 2019.

[60] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. CVPR*, 2015, pp. 787–796.

[61] J. Cao, Y. Hu, B. Yu, R. He, and Z. Sun, "Load balanced GANs for multi-view face image synthesis," 2018, *arXiv:1802.07447*. [Online]. Available: http://arxiv.org/abs/1802.07447

[62] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proc. ICCV*, 2017, pp. 3990–3999.

[63] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. NIPS*, 2015, pp. 2017–2025.

[64] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, and F. Zhao, "Towards pose invariant face recognition in the wild," in *Proc. CVPR*, 2018, pp. 2207–2216.

[65] J. Zhao *et al.*, "3D-aided deep pose-invariant face recognition," in *Proc. IJCAI*, 2018, pp. 1184–1190.

[66] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Learning a high fidelity pose invariant model for high-resolution face frontalization," in *Proc. NIPS*, 2018, pp. 2867–2877.

[67] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou, "Uv-GAN: Adversarial facial UV map completion for pose-invariant face recognition," in *Proc. CVPR*, 2018, pp. 7093–7102.

[68] A. Fawzi, H. Samulowitz, D. Turaga, and P. Frossard, "Image inpainting through neural networks hallucinations," in *Proc. Image, Video, Multidimensional Signal Process. Workshop*, 2016, pp. 1–6.

[69] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proc. CVPR*, 2017, pp. 5485–5493.

[70] 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.

[71] J. Johnson, A. Alahi, and F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016.

[72] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, May 2010.

[73] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep., 2007.

[74] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

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

[76] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *Proc. CVPR*, 2015, pp. 676–684.

[77] J. Xu, A. Ghodrati, M. Pedersoli, and T. Tuytelaars, "Towards automatic image editing: Learning to see another you," in *Proc. BMVC*, 2016.

[78] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *Proc. ICCV*, 2013, pp. 113–120.

[79] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: A deep model for learning face identity and view representations," in *Proc. NIPS*, 2014, pp. 217–225.

Qingyan Duan (Student Member, IEEE) graduated from the Hefei University of Technology, Hefei, China, in 2012. She received the M.Sc. degree from Chongqing University, Chongqing, China, in 2016, where she is currently pursuing the Ph.D. degree.

Her current research interests include deep learning, pattern recognition, and computer vision.

Lei Zhang (Senior Member, IEEE) received the Ph.D. degree in circuits and systems from the College of Communication Engineering, Chongqing University, Chongqing, China, in 2013.

He was a Post-Doctoral Fellow with The Hong Kong Polytechnic University, Hong Kong, from 2013 to 2015. He is currently a Professor/ Distinguished Research Fellow with Chongqing University. He has authored more than 100 scientific articles in top journals, such as the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and top conferences, such as ICCV, CVPR, AAAI, ACM MM, and ACCV. His current research interests include machine learning, pattern recognition, computer vision, and intelligent systems.

Dr. Zhang was a recipient of the Best Paper Award of CCBR2017, the Outstanding Reviewer Award of many journals, such as *Pattern Recognition* and *Information Sciences*, the Outstanding Doctoral Dissertation Award of Chongqing, China, in 2015, the Hong Kong Scholar Award in 2014, the Academy Award for Youth Innovation in 2013, and the New Academic Researcher Award for Doctoral Candidates from the Ministry of Education, China, in 2012. He serves as an Associate Editor for the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, *Neural Networks* (Elsevier), and so on.