Face De-occlusion using 3D Morphable Model and Generative Adversarial Network

Xiaowei Yuan and In Kyu Park
{xiaoweichn@qq.com pik@inha.ac.kr}
Dept. of Information and Communication Engineering, Inha University, Incheon 22212, Korea

Abstract

In recent decades, 3D morphable model (3DMM) has been commonly used in image-based photorealistic 3D face reconstruction. However, face images are often corrupted by serious occlusion by non-face objects including eyeglasses, masks, and hands. Such objects block the correct capture of landmarks and shading information. Therefore, the reconstructed 3D face model is hardly reusable. In this paper, a novel method is proposed to restore de-occluded face images based on inverse use of 3DMM and generative adversarial network. We utilize the 3DMM prior to the proposed adversarial network and combine a global and local adversarial convolutional neural network to learn face de-occlusion model. The 3DMM serves not only as geometric prior but also proposes the face region for the local discriminator. Experiment results confirm the effectiveness and robustness of the proposed algorithm in removing challenging types of occlusions with various head poses and illumination. Furthermore, the proposed method reconstructs the correct 3D face model with de-occluded textures.

1. Introduction

3D face reconstruction from a single image is a key technology in many computer vision and graphic applications, such as face recognition and face animation. Since Blanz and Vetter [1] proposed the 3D morphable face model (3DMM), the methodology based on 3DMM has been most popular for coarse face geometric reconstruction. Further development involves using the shape from shading (SfS) technique to enhance details (*e.g.* wrinkles) on the face geometry [14, 22, 13]. These techniques assume that the input image is free from occlusion, or at most, self-occluded by head pose variation. However, in actual situations in the wild, we encounter new challenges in which existing algorithms become inapplicable due to serious occlusion by eyeglasses, masks, hands, and others.

To solve this problem in face recognition, a few methods propose solutions for automatic face de-occlusion to improve recognition performance [4, 21, 27, 31, 18]. However, almost all existing methods work under highly constrained conditions, *e.g.* low-resolution grayscale images with predefined head pose. Therefore, these methods cannot perform photorealistic 3D face reconstruction in terms of image resolution and image diversity.

(a) (b) (c) (d)

Figure 1: Face de-occlusion results by applying the proposed method. (a)(b) Real images. (c)(d) Synthetic images. (Upper) Input images with occlusions. (Lower) De-occlusion results.

The recent work by Tran [26] addresses the challenge of detailed face reconstruction from occluded images. However, rather than performing de-occlusion, this method focuses on geometrical reconstruction by searching the reference dataset to reconstruct the bump map on the occluded region. By contrast, our proposed method directly removes the occlusion on the face image, thereby allowing texture mapping on the reconstructed 3D model.

In this paper, we address the problem of face de-occlusion from seriously occluded actual images while focusing on the application of 3D face reconstruction. The proposed method aims to directly remove occlusions, thereby enabling it to synthesize the texture for the 3D face model. An example of face de-occlusion is shown in Figure 1. Various occlusions and head poses are observed from actual face images. Thus, automatically removing face occlusions through a purely data-driven manner is a challeng-

ing task.

To solve this problem, we propose a novel 3DMM-conditioned deep convolutional neural network to learn to automatically remove occlusions. To the best of our knowledge, this is the first face de-occlusion network that attempts to exploit the potential use of 3DMM for face de-occlusion. Similar with the previous works [5, 17, 29], we employ the recent generative adversarial network (GAN) [10] that has been widely used to train an image synthesis model with strong ability to generate natural and high-quality images. In the proposed approach, global and local discriminators are combined with the generative model to achieve high-quality image synthesis. During training, 3DMM not only serves as prior but also proposes face region for the local discriminator. To diversify the training images with various occlusions, we synthesize a large-scale dataset from 300W-3D and AFLW2000-3D [32]. The key contributions of this paper can be summarized as follows:

- We propose a novel deep face de-occlusion framework, which applies the inverse use of 3DMM and GAN and consists of a generator and two discriminators.
- The proposed face de-occlusion model can handle face images under challenging conditions, *e.g.*, serious occlusions with nontrivial head poses and illumination variations.
- We build a large-scale synthesized face-with-occlusion dataset. All occlusions are semantically placed on the face with reference to face landmarks.
- The proposed face de-occlusion method not only boosts the performance of 3D face reconstruction but also allows face attribute editing by modifying the 3DMM coefficients.

2. Related Works

Image Completion Image completion or image inpainting aims to recover masked or missing regions on images with visually plausible contents. Recently, the generative model has been widely used in image completion with reasonably acceptable results [11, 5, 17, 25, 30]. These methods train an auto-encoder to predict the missing region by using a combination of reconstruction loss and adversarial loss. Despite the ability of the image completion technique to recover high-quality visual patterns in face de-occlusion tasks, the occluded region needs to be masked manually or with an additional object detection algorithm to segment the occluder. By contrast, the proposed model does not need any preprocessing on the occluded region and can automatically remove the occlusion.

Face De-occlusion and Frontalization Conventional face de-occlusion algorithms are developed to increase the performance of face recognition algorithms. Wright *et al.* [27] proposed to apply sparse representation to encode faces and demonstrated the robustness of the extracted features to occlusion. Cheng *et al.* [4] introduced the double-channel SSDA (DC-SSDA) to detect noise by exploiting the difference between activations of two channels. Recently, a deep learning-based approach has been proposed by Zhao [31] to restore the partially occluded face in several successive processes using an LSTM auto-encoder.

However, these methods can only remove occlusions under constrained conditions. Images in low-resolution grayscale and all faces in the dataset have to be cropped and aligned first. Therefore, these methods cannot conduct practical applications beyond face recognition. By contrast, the proposed method is targeted to perform actual face reconstruction and texture synthesis. Thus, we generalize the input to RGB images with enlarged resolution (256×256) and with various head poses.

Besides, Yin *et al*. proposed a deep 3DMM-conditioned face frontalization method called FF–GAN [29], which incorporates 3DMM coefficients into the GAN structure to provide poses prior to the face frontalization task. This method utilizes 3DMM coefficients as a weak prior to reduce the artifacts during frontalization in extreme profile views.

3D Face Reconstruction from Occluded Image Bernhard *et al.* [6] proposed an occlusion-aware face modeling method in which they incorporated 3DMM as appearance prior in a RANSAC-like algorithm. In this method, the input image is segmented into face and non-face regions and the illumination is estimated using the face region only. However, this method is not robust if the occlusions, e.g., hands, have a similar color appearance to the face. A recent face alignment technique shows the robustness in occluded face images [2, 28]. Thus, this technique can be used to fit 3DMM and produce a 3D face model with a few details. Although their goal is to robustly find the head pose, de-occlusion in not within the scope of their work. Tran *et al.* [26] is the first to address the problem of detailed face reconstruction from occluded images by filling in the corrupted region of the bump map using a similar patch in a reference dataset. Although this method can generate a complete representation of face details, the de-occluded face image is not reconstructed [26].

Face Synthesis with GAN GAN [10] utilizes min–max optimization over the generator and discriminator and shows significant improvement in face synthesis applications, such as face attribute editing [24], and face completion [5, 17]. Gecer *et al.* exploited to synthesize facial images [8] and facial textures [9] conditioned on latent 3DMM parameters. However, no previous study has been conducted on using GAN for de-occlusion on challenging faces.

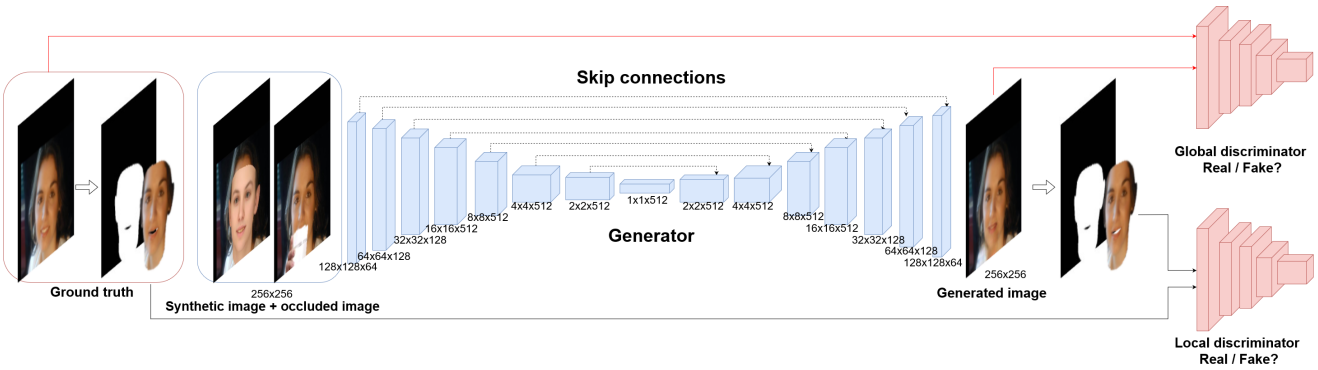


Figure 2: Proposed network structure. It consists of a generator with two discriminators. The generator takes a synthesis image and an occluded image as input. Two discriminators help to generate a more natural result. Only the generator is necessary during testing.

3. Overview and Background

3.1. Overview of the Proposed Method

In this paper, we address the face de-occlusion problem by incorporating 3DMM and GAN in the same framework. Motivated by [29], we propose to use 3DMM as our geometric regularization in our face de-occlusion model. In our work, 3DMM is a strong prior without which the algorithm would fail completely. That is, 3DMM is used to provide constraints on the appearance of the occluded region in which the generator output is used explicitly to synthesize the de-occluded image.

We first fit the 3DMM to an occluded image and synthesize a 2D face image. Then, we take the synthesis and occluded face as the inputs of the generator to synthesize occlusion-free images. At the same time, a global discriminator and a local discriminator attempt to distinguish the image as a real image or a generated one. The 3DMM serves not only as the geometric prior but also provides the face region for the local discriminator. With the guidance of 3DMM, the generator can efficiently remove occlusions even on challenging faces. Figure 2 illustrates our proposed framework consisting of a 3DMM-conditioned generator and two discriminators.

3.2. 3D Morphable Model

3DMM is the most commonly used statistical method for the representation and synthesis of face geometry and texture. In our work, we use a multilinear 3DMM with 53K vertices and 106K triangles to represent the 3D face shape [33]. Each face geometry can be parameterized as follows:

$$M(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \bar{\boldsymbol{S}}_{\boldsymbol{id}} + \boldsymbol{\alpha} \cdot \boldsymbol{S}_{\boldsymbol{id}} + \boldsymbol{\beta} \cdot \boldsymbol{S}_{\boldsymbol{exp}}, \tag{1}$$

3DMM assumes that each face shares a similar structure that distributes around the average identity $\bar{\boldsymbol{S}}_{\boldsymbol{id}} \in R^{3n}$. $\boldsymbol{S}_{\boldsymbol{id}} \in R^{3n \times 80}$, $\boldsymbol{S}_{\boldsymbol{exp}} \in R^{3n \times 29}$ are principal components representing the basis of identity and expression. $\boldsymbol{\alpha} \in R^{80}$ and $\boldsymbol{\beta} \in R^{29}$ are the use-specific coefficients estimated from the given image. In our implementation, the identity component comes from the Basel Face Model (BFW) [1], whereas the expression comes from the FaceWarehouse database [3].

Synthesis is dependent on the 3DMM coefficients $\boldsymbol{\alpha}, \boldsymbol{\beta}$, the rigid translation $\boldsymbol{R}, \boldsymbol{t}$, and the camera projection matrix $\boldsymbol{\Pi}$. To reconstruct a 3D face model, we align corresponding 2D face landmarks with 3D landmarks on the bilinear face model using the pose normalization method [15]. All 3DMM parameters, defined as $\boldsymbol{\Theta}$, are then jointly estimated by using the following formula:

$$\arg\min_{\boldsymbol{\Theta}} = \|\boldsymbol{\Pi}(\boldsymbol{RV} + \boldsymbol{t}) - \boldsymbol{U}\|^2 + \rho_1 \|\frac{\boldsymbol{\alpha}}{\boldsymbol{\xi_{id}}}\|^2 + \rho_2 \|\frac{\boldsymbol{\beta}}{\boldsymbol{\xi_{exp}}}\|^2, \tag{2}$$

where $\boldsymbol{U}$ represents 2D face landmarks, and $\boldsymbol{V}$ represents corresponding vertices on the face model determined by 3DMM coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$. ρ_1 and ρ_2 are positive weights of the regularization term to enforce the parameters to stay statistically close to the mean. $\boldsymbol{\xi_{id}}$ and $\boldsymbol{\xi_{exp}}$ are the standard deviations of shape and expression basis, respectively. We employ an occlusion-robust face alignment method [2] to infer 68 face landmarks. All parameters are jointly solved via the Levenberg–Marquart algorithm [20].

Based on the fitting result, the synthesis is generated as shown in Figure 3. The correspondence between pixels and triangles is computed by using Z buffering. Finally, the 3DMM is occlusion-free and can synthesize the appearance and pose of a face.

4. Face De-occlusion using GAN

The main framework of our model is a GAN that consists of a generator G, a global discriminator D_g, and a local

Figure 3: Occluded face image (left) and the 3DMM synthesis (right).

discriminator D_l. The generator takes the occluded image and 3DMM synthesis as input to generate the occlusion-free image. Moreover, two discriminators D_g and D_l attempt to determine whether the generator output is a real face image or not. 3DMM not only serves as the prior but also provides a mask indicating the face region for the local discriminator. Additionally, a smoothness term is used to regularize G to generate an image with fewer artifacts.

4.1. Generator Module

The generator G works as an auto encoder–decoder to remove face occlusion and construct the corrupted region. The occluded image $\boldsymbol{I}$, concatenated with the synthesis $\boldsymbol{I^s}$, is first mapped into the hidden feature through the encoder, which captures not only the variation of the known region but also the coarse geometric information of the occluded region. Then, the feature vector is fed into a decoder to generate an occlusion-free image.

Our encoder and decoder use modules of the form Convolution–BatchNorm –Relu and have the same architecture except for the input layer. We follow the encoder–decoder network designed in [12], where a skip layer is used to preserve the low-level feature from the corresponding symmetrical layer. The skip collection allows combining the coarse geometry information from the downsampling path with the high-frequency features in the upsampling path to finally generate an occlusion-free image with good visual quality.

Even though 3DMM can synthesize the appearance and pose of a face, the generated image looks unrealistic and tends to lose all face details. To force the generator to output photo-realistic images, we adopt a pixel-wise L_1 reconstruction loss to penalize the output from the ground truth by using the following equation:

$$L_{gen} = |G(\boldsymbol{I}, \boldsymbol{I^s}) - \boldsymbol{I^g}|_1, \tag{3}$$

where $\boldsymbol{I^g}$ is the ground truth, $\boldsymbol{I}$ is the occluded image, and $\boldsymbol{I^s}$ is the synthesis of 3DMM. As actual face images have various head poses, we avoid using symmetry loss as in the work of Yin [29].

Despite the ability of the generator to reconstruct the occluded region with semantical contents, inconsistency occurs especially when the occlusion has a complex pattern. Thus, we use a total variation regularization to reduce the artifacts on the reconstructed region. We perform a L_2 minimization to the gradient of the generated image. The regularization is performed separately for each coordinate and then combined. The total variation regularization is commonly used in image noise removal using the following equation:

$$L_{tv} = |\nabla_x G(\boldsymbol{I}, \boldsymbol{I^s})|_2 + |\nabla_y G(\boldsymbol{I}, \boldsymbol{I^s})|_2 \tag{4}$$

However, this term tends to smoothen high-frequency details. Thus, we multiply a small weight to L_{tv} to avoid oversmoothening.

Our generator can remove the occlusion and generate photo-realistic contents. However, recovering the expression in the occluded region is an ill-posed problem. Expression coefficients estimated from the occluded region can be arbitrary. Our generator relies on 3DMM for geometric information. Thus, we can edit face attributes by simply adjusting the 3DMM coefficients. Therefore, face deocclusion and face attributes are integrated into one framework. The generator output of the occlusion-free image is consistent with the synthesis of 3DMM in geometry and demonstrates the solid effect of the 3DMM in regularizing the generation process.

4.2. Discriminator Module

Reconstruction loss tends to average all the details, thereby making the synthesized contents look blurry. Moreover, the generator only optimizes on the occluded region and cannot learn the relationship between pixels, which results in the generated contents being discontinuous with surroundings.

Recently, GAN consisting of generator and discriminator has been widely used for image synthesis. In this work, the generator synthesizes an occlusion-free face image, whereas the discriminator determines whether the generated face is real or not. The min-max optimization over generator and discriminator forces the model to synthesize images with better visual quality.

Our discriminator includes a local discriminator D_l and a global discriminator D_g. The latter is used to determine the faithfulness of the entire image to enforce the generated region to become consistent with the surroundings. Considering our goal to reconstruct face geometry and texture synthesis on the image, we only rely on the face region. Thus, we enforce the optimization of the local discriminator in the face region. The mask $\mathcal{M}$ used for the local discriminator is a projected silhouette from the 3DMM indicating the face region. Compared with the global discriminator,

the local module enhances details in the face region with well-defined boundaries and less noise.

By combining the local module with the global module, we not only guarantee the statistical consistency of the generated face region with its surroundings but also encourage the recovered face region to become highly informative. To train these two discriminators, the following objectives are minimized:

$$\begin{aligned} L_{D_g} = &- \mathbb{E}_{\boldsymbol{I^g} \in R} \log D_g(\boldsymbol{I^g}) \\ &- \mathbb{E}_{\boldsymbol{I^s} \in K} \log(1 - D_g(G(\boldsymbol{I}, \boldsymbol{I^s}))), \end{aligned} \tag{5}$$

$$\begin{aligned} L_{D_l} = &- \mathbb{E}_{\boldsymbol{I^g} \in R} \log D_l(\mathcal{M} \odot (\boldsymbol{I^g})) \\ &- \mathbb{E}_{\boldsymbol{I^s} \in K} \log(1 - D_l(\mathcal{M} \odot G(\boldsymbol{I}, \boldsymbol{I^s}))), \end{aligned} \tag{6}$$

where the $\odot$ denotes the element-wise multiplication and R and K are real and generated image sets, respectively. Our two discriminators have similar network structures that consist of seven convolution layers. After the last layer, a convolution is mapped to one-dimensional output, followed by a sigmoid function. The outputs of the discriminators determine whether the probability of the input is real or generated.

In addition, G attempts to fool the two discriminators in identifying the generated image as real by minimizing the following loss:

$$\begin{aligned} L_{adv_l} &= - \mathbb{E}_{\boldsymbol{I^s} \in K} \log(D_l(\mathcal{M} \odot G(\boldsymbol{I}, \boldsymbol{I^s}))) \\ L_{adv_g} &= - \mathbb{E}_{\boldsymbol{I^s} \in K} \log(D_g(G(\boldsymbol{I}, \boldsymbol{I^s}))) \end{aligned} \tag{7}$$

4.3. Objective Function

To summarize, the final loss for our proposed 3DMM-conditioned GAN is represented as a weighted sum of the aforementioned losses:

$$\begin{aligned} L = &\lambda_1 L_{gen} + \lambda_2 L_{tv} + \lambda_3 L_{adv_g} + \\ &\lambda_4 L_{adv_l} + \lambda_5 L_{D_l} + \lambda_6 L_{D_g} \end{aligned} \tag{8}$$

Weights λ_1, λ_2, λ_3, λ_4, λ_5, and λ_6 are used to balance different terms.

5. Experimental Result

5.1. Dataset

Occluded images, 3DMM synthesis, and corresponding occlusion-free images are needed to train our face de-occlusion model. Owing to the difficulty in collecting sufficient occluded face images with their corresponding occlusion-free images, we train our model on our synthesized dataset. The datasets used for training and testing are introduced in the following.

300W-3D This dataset consists of 7,700 300-W [23] samples with the fitted 3DMM parameters and 68 face landmarks for each sample. All images in 300-W are real photos and cover variations in pose, illumination, background, and image quality.

Figure 4: Samples of our synthesized dataset for training. Occlusions are located semantically based on the face landmarks.

AFLW2000-3D This dataset consists of the first 2,000 images in AFLW [16]. Similar to 300W-3D, 3DMM parameters and 68 landmarks are provided for each image.

CelebA [19] This dataset consists of 202,599 celebrity images with each image cropped and roughly aligned by the positions of the two eyes. We select occluded faces in this dataset for testing.

We synthesize occlusions caused by six common objects on occlusion-free faces in 300W-3D and AFLW2000-3D. These objects include masks, eyeglasses, sunglasses, cups, scarves, and hands. We layer these occlusions on the specific location of the face with reference to the face landmarks. Figure 4 shows examples of the occluded faces generated using this approach. All occlusions are semantically located on the face to augment the reality of our dataset. Then, we generate the synthesized image of 3DMM for every training sample by using 3DMM coefficients and camera pose provided by 300W-3D and AFLW2000-3D. We synthesize a dataset with a total of 134,233 occluded images. All faces in the dataset are resized 256×256 and with head poses varying from 60° to 60°. We select 132,233 images for training and 2,000 images for testing. Random cropping and horizontal flipping are used in data augmentation to avoid overfitting. Besides using the synthesized images, we test our model on real images, which consist of the occluded images from the aforementioned three datasets.

5.2. Implementation Details

We train our network with batch size 5 and utilize Adam optimizer. Instead of jointly training all modules, we gradually add them. In the first stage, we train the generator and global discriminator with a learning rate of 0.0002 for 100 epochs. In the second stage, we add the local discriminator and remove the total variation regularization to finetune the network with a learning rate of 0.00005, and train another 10 epochs. During training, we set the value of λ_1, λ_2, λ_3, λ_4, λ_5, λ_6 as 10, 10^{-5}, 1, 1, 1, and 1, respectively. In the testing stage, only the generation module is required. The

Figure 5: Face de-occlusion results on the synthetic dataset. (a) Image with occlusion. (b) De-occlusion result. (c) Original image (ground truth).

Figure 6: Face de-occlusion results on the real dataset. (a) Image with occlusion. (b) De-occlusion result.

entire training procedure takes approximately 4d on a single GeForce GTX 1080Ti GPU. In the testing stage, a 256×256 color image can be processed in under a second.

5.3. Face De-occlusion

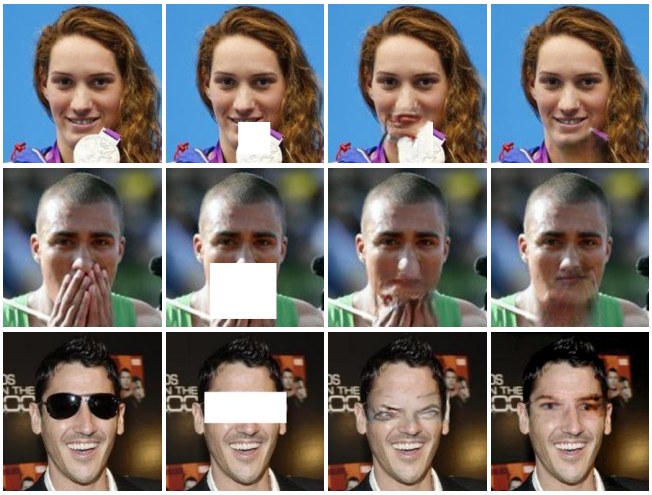
Qualitative Result Figure 5 and Figure 6 show the face de-occlusion results on the synthetic and real images, respectively. Note that the identities in the test dataset are separated from the training dataset. As shown in Figure 5(b) and Figure 6(a), test images have various types of occlusion at arbitrary locations. The results show that the proposed method successfully removes the occlusion and generates a photorealistic de-occluded image for both synthetic and real data even when a significant portion of the face region is occluded. As shown in the last two examples in Figure 6, the proposed method model removes not only occlusions similar to that in our training dataset but also those that do not exist in our dataset (no occlusion with microphones in the training dataset). The result confirms that the proposed 3DMM-conditioned face de-occlusion model can remove different types of occlusions with challenging conditions including various head poses and illumination.

Failure occurs when more than one type of occlusion exists on the face and when the occlusion is located out of the synthesis range of the 3DMM parametric space, *e.g.*, hands above the forehead.

Type of occlusion	PSNR	SSIM
Lower face	27.3228	0.9615
Upper face	34.0024	0.9860
Left/right half of face	28.7785	0.9659
Three quarters of face	22.1680	0.8967

Table 1: Quantitative evaluation for different types of occlusion.

Quantitative Result To quantitatively measure the de-occlusion performance, two popular metrics, *i.e.*, PSNR and SSIM, are evaluated on the de-occlusion result of the syn-

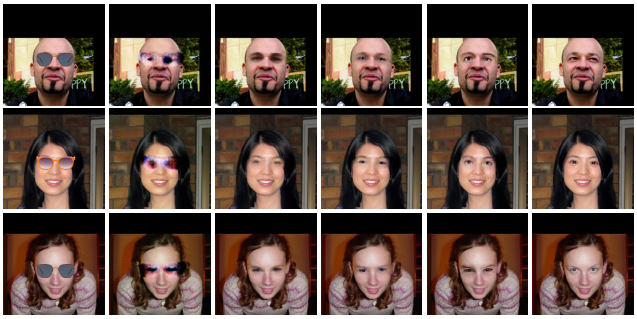


(a) (b) (c) (d)

Figure 7: Comparison result with the state-of-the-art face inpainting algorithm [30]. (a) Input image with occlusion. (b) Masking the occluded region. (c) De-occlusion result with the inpainting algorithm. (d) De-occlusion result with the proposed algorithm.

thetic dataset and listed in Table 1. The performance of our model slightly drops when more than half of the face is occluded, which is expected as a large occlusion size indicates uncertainty in pixel values. The model also shows better de-occlusion performance on the upper face than the lower face because occlusions on the lower face region have complex patterns, such as different scarves and cups.

Comparison The goal of the proposed face de-occlusion model is to remove occlusions on face images and recover the missing region. As the goal of the previous methods [27, 4, 31] is completely different (face recognition with a low-resolution grayscale image), we do not compare our results with theirs. Instead, we compare our approach with the recent state-of-art face inpainting method [30] because the recent face inpainting method shows potential application in removing occlusions and reconstructing de-occluded face regions. As the method proposed by Yu [30] is only trained on the *CelebA* dataset, we conduct the experiment on occluded images from that dataset to be fair. First, the occluded region is masked with the provided pattern and the inpainting algorithm is applied to reconstruct the masked region. As shown in Figure 7, the inpainting algorithm does not work effectively on face images. Face inpainting is usually utilized on a well-aligned dataset, thereby failing to generate the semantic contents on difficult cases, such as posed face and complex backgrounds. On the contrary, the proposed method can automatically remove occlusions without any preprocessing on the occluded region



(a) (b) (c) (d) (e) (f)

Figure 8: Comparison result under different setting. (a) Occluded face image. (b) Without 3DMM synthesis. (c) With generator only. (d) With global discriminator. (e) With global and local discriminators. (f) Ground truth.

while showing significantly better results.

Ablation Study To validate the effects of the 3DMM synthesis, we train the other variants with similar hyperparameters but different settings and compare the performance. We remove 3DMM, global discriminator, and local discriminator in turn. Without 3DMM, the network generates noisy outputs or fails to generate informative results. The result is sensible because generator with only pixel-wise reconstruction loss is too weak to learn the representation of the face geometry from a challenging face dataset. Note that, in face de-occlusion, we have to find and restore the occluded region while handling the pose variation simultaneously which is a serious ill-posed problem. By using 3DMM as a prior, the ill-posedness can be alleviated and de-occlusion on images with various head poses can be performed properly. Without discriminators, the model can generate images with semantical contents but artifacts remain on the recovered region. With only the global discriminator, the result looks sharp and coherent, but lacks details in the eyes. With the combined global and local discriminators, the face de-occlusion results look visually realistic. The visual comparison results are summarized in Figure 8.

5.4. 3D Face Reconstruction

As our motivation is face de-occlusion for 3D face reconstruction, we conduct the experiments to investigate the effect of our face de-occlusion model on 3D face reconstruction. In our experiment, coarse 3D face model and detailed face geometry are reconstructed with the de-occluded face image. For this purpose, conventional landmark-based 3DMM fitting is conducted first and the shape-from-shading (SfS) method is employed to enhance details on the coarse face model [14, 22, 13]. Based on the assumption that Lambertian reflection on the face exists, the intensity

(a) (b) (c) (d) (e)

Figure 9: Comparison results of detailed face reconstruction from occluded image and de-occluded image. (a) Occluded face image. (b) Face reconstruction result with (a). (c) De-occluded image. (d) Face reconstruction result with (c). (e) 3D face model with de-occluded texture mapping.

formation of the face image can be represented as follows:

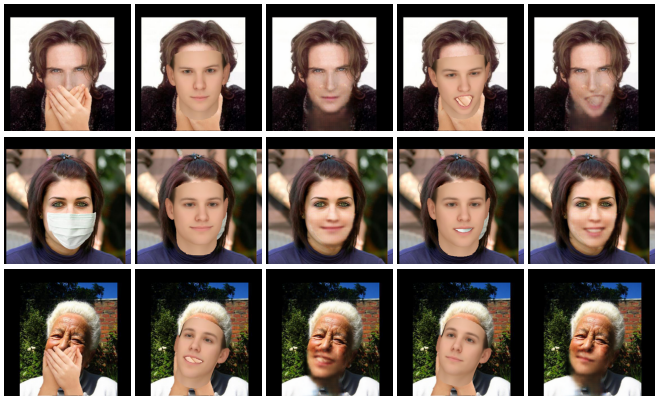
$$\boldsymbol{I}(x,y) = \boldsymbol{\rho}(x,y)\vec{\boldsymbol{l}}\boldsymbol{Y}(\vec{\boldsymbol{n}}(x,y)), \tag{9}$$

where $\boldsymbol{Y}(\vec{\boldsymbol{n}}(x,y))$ is the second-order spherical harmonics [7], $\vec{\boldsymbol{l}}$ represents lighting coefficients, $\boldsymbol{\rho}(x,y)$, and $\vec{\boldsymbol{n}}(x,y)$ are the albedo and normal vector at pixel(x,y), respectively. Following the work of Kemelmacher [14], we estimate lighting $\vec{\boldsymbol{l}}$, albedo $\boldsymbol{\rho}(x,y)$, and normal vector $\vec{\boldsymbol{n}}(x,y)$ in turn. Then, the estimated normal vector $\vec{\boldsymbol{n}}(x,y)$ is integrated to recover the detailed face geometry.

Figure 9 shows the comparison results of detailed 3D face reconstruction using face images with occlusion and de-occlusion. The results demonstrate that occlusion causes significant noise on the face geometry. Note that the geometry is not completely corrupted because the shape is still controlled by 3DMM. On the contrary, by removing occlusions using the proposed method, both face geometry and textured 3D face model are reconstructed correctly.

5.5. Face Attributes Manipulation

The proposed face de-occlusion can be applied to future studies, such as face editing and face recognition, to improve performance. In this further experiment, we show the application of the proposed model in face attribute editing.



(a) (b) (c) (d) (e)

Figure 10: Results of face attribute editing. (a) Occluded face image. (b)(d) 3DMM synthesis with different expression coefficients. (c) Generated face guided by (b). (e) Generated face guided by (d).

As the proposed generative model recovers the occluded face guided by the 3DMM synthesis, it allows face attribute editing by simply adjusting the 3DMM coefficients to any desired one. Therefore, the proposed model holds potential application in face editing to generate a novel portrait, as shown in Figure 10. Given the same occluded image, we can modify the attribute of the generated face by changing 3DMM expression coefficients as shown in Figure 10(b) and (d).

6. Conclusion

In this paper, we proposed a 3DMM-conditioned GAN framework to remove face occlusion and restore the occluded region. To the best of our knowledge, this study is the first to explore the use of 3DMM in face de-occlusion on challenging dataset. Experimental results show that our face de-occlusion model can remove face occlusion on synthetic and real images. The proposed method not only removes the occlusion but also reconstructs the correct 3D face model without occluded texture. Furthermore, our method allows face attribute editing by simply modifying the 3DMM coefficients.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019R1A2C1006706).

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proc. of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999. 1, 3

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D and 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *Proc. IEEE International Conference on Computer Vision*, 2017. 2, 3

[3] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. FaceWarehouse: A 3D facial expression database for visual computing. *IEEE Trans. on Visualization and Computer Graphics*, 20(3):413–425, 2014. 3

[4] Lele Cheng, Jinjun Wang, Yihong Gong, and Qiqi Hou. Robust deep auto-encoder for occluded face recognition. In *Proc. ACM International Conference on Multimedia*, pages 1099–1102, 2015. 1, 2, 7

[5] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. UV-GAN: Adversarial facial uv map completion for pose-invariant face recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[6] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3D morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126(12):1269–1287, 2018. 2

[7] Darya Frolova, Denis Simakov, and Ronen Basri. Accuracy of spherical harmonic approximations for images of lambertian objects under far and near lighting. In *Proc. European Conference Computer Vision*, pages 574–587, 2004. 8

[8] Baris Gecer, Binod Bhattarai, Josef Kittler, and Tae-Kyun Kim. Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3D morphable model. In *Proc. European Conference Computer Vision*, pages 230–248, October 2018. 2

[9] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3D face reconstruction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1164, June 2019. 2

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Advances in Neural Information Processing Systems*, pages 2672–2680. 2014. 2

[11] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Trans. on Graphics.*, 36(4):107:1–107:14, 2017. 2

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 4

[13] Luo Jiang, Juyong Zhang, Bailin Deng, Hao Li, and Ligang Liu. 3D face reconstruction with geometry details from a single image. *IEEE Trans. on Image Processing*, 27(10):4756–4770, 2018. 1, 7

[14] Ira Kemelmacher-Shlizerman and Ronen Basri. 3D face reconstruction from a single image using a single reference face shape. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(2):394–405, 2011. 1, 7, 8

[15] Ira Kemelmacher-Shlizerman and Steven M. Seitz. Face reconstruction in the wild. In *Proc. International Conference on Computer Vision*, pages 1746–1753, 2011. 3

[16] Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Proc. IEEE International Conference on Computer Vision Workshops*, pages 2144–2151, 2011. 5

[17] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[18] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proc. European Conference Computer Vision*, pages 89–105, 2018. 1

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. The IEEE International Conference on Computer Vision*, 2015. 5

[20] Jorge J. Moré. The levenberg-marquardt algorithm: Implementation and theory. In *Numerical Analysis*, pages 105–116, 1978. 3

[21] Jeong-Seon Park, You Hwa Oh, Sang Chul Ahn, and Seong-Whan Lee. Glasses removal from facial image using recursive error compensation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27:805–811, 2005. 1

[22] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 7

[23] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013. 5

[24] Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[25] Linsen Song, Jie Cao, Linxiao Song, Yibo Hu, and Ran He. Geometry-aware face completion and editing. *CoRR*, abs/1809.02967, 2018. 2

[26] Anh T. Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3D face reconstruction: Seeing through occlusions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2

[27] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(2), 2009. 1, 2, 7

[28] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 2

[29] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *Proc. IEEE International Conference on Computer Vision*, 2017. 2, 3, 4

[30] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 7

[31] Fang Zhao, Jiashi Feng, Jian Zhao, Wenhan Yang, and Shuicheng Yan. Robust lstm-autoencoders for face deocclusion in the wild. *IEEE Trans. on Image Processing*, 27(2):778–790, 2018. 1, 2, 7

[32] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3D solution. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[33] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015. 3