Complete Face Recovery GAN: Unsupervised Joint Face Rotation and De-Occlusion from a Single-View Image

Yeong-Joon Ju\textsuperscript{1*} Gun-Hee Lee\textsuperscript{2*} Jung-Ho Hong\textsuperscript{1} Seong-Whan Lee\textsuperscript{1†}
\textsuperscript{1}Department of Artificial Intelligence, Korea University, Seoul, South Korea
\textsuperscript{2}Department of Computer and Radio Communications Engineering, Korea University, Seoul, South Korea
\{yj_ju, gunhlee, jungho-hong, sw.lee\}@korea.ac.kr

Abstract

Although various face-related tasks have significantly advanced in recent years, occlusion and extreme pose still impede the achievement of higher performance. Existing face rotation or de-occlusion methods only have emphasized the aspect of each problem. In addition, the lack of high-quality paired data remains an obstacle for both methods. In this work, we present a self-supervision strategy called Swap-R&R to overcome the lack of ground-truth in a fully unsupervised manner for joint face rotation and de-occlusion. To generate an input pair for self-supervision, we transfer the occlusion from a face in an image to an estimated 3D face and create a damaged face image, as if rotated from a different pose by rotating twice with the roughly de-occluded face. Furthermore, we propose Complete Face Recovery GAN (CFR-GAN) to restore the collapsed textures and disappeared occlusion areas by leveraging the structural and textural differences between two rendered images. Unlike previous works, which have selected occlusion-free images to obtain ground-truths, our approach does not require human intervention and paired data. We show that our proposed method can generate a de-occluded frontal face image from an occluded profile face image. Moreover, extensive experiments demonstrate that our approach can boost the performance of facial recognition and facial expression recognition. The code is publicly available\textsuperscript{1}.

1. Introduction

Various studies have been conducted on face-related tasks, including facial recognition, expression recognition, and re-identification with progress in a deep neural network. Despite recent improvements, extreme pose and occlusion remain obstacles to the above tasks. Face rotation and de-

\textsuperscript{*}Equal contribution. \textsuperscript{†}Corresponding author.
\textsuperscript{1}https://github.com/yeongjoonJu/CFR-GAN.

Figure 1: Qualitative results on CelebA-HQ and FFHQ datasets. Our method is able to synthesize photorealistic rotated and de-occluded face images, achieving the state-of-the-art performance on standard benchmarks.

occlusion can alleviate these problems but are challenging tasks because of the lack of high-quality training data.

Most traditional methods for face rotation use the 3D Morphable Model (3DMM) \cite{Blanz99}, a statistical model of facial shape and texture that uses a set of linear basis functions \cite{Blanz99, Weinfurter10, Klassen15}. A challenging issue is the natural estimation of the texture of the invisible face area. Zhu et al. \cite{Zhu18} proposed symmetric editing and invisible region filling to solve this problem. However, these methods tend to show unnatural results with visible artifacts. Recently, many studies \cite{Zhu18, Liu20, Wei20} have been proposed to synthesize photorealistic rotated faces by utilizing the power of the Generative Adversarial Network (GAN) \cite{Goodfellow14}. These methods have shown remarkable performance improvements but often lose the local facial details of the face. In addition, they do not generalize well beyond the controlled dataset for training, and the resulting images are usually limited to low resolution, which is not perceptually satisfactory.

The lack of high-quality paired training data is also a critical issue in de-occlusion tasks. Existing methods \cite{Zhu18, Liu20, Wei20} artificially synthesize images to occlude the parts of the face, training a deep neural network to restore the original face images from unnaturally synthesized images. However, since these methods depend on artificially synthesized data, they tend to show unnatural results for various
occlusions. To address these limitations, we propose Complete Face Recovery GAN (CFR-GAN), a fully unsupervised method for joint face rotation and de-occlusion. Our method covers two challenging tasks: (i) estimating the mask for the occlusion area that can provide 3D face-based guidance to naturally restore the texture of the occlusion area, and (ii) providing strong self-supervision for joint face rotation and de-occlusion by proposing a Swap-R&R strategy that transfers occlusions from an image to the estimated 3D face and rotates it twice, as if rotated from a different pose.

First, two 3D faces are generated from the input image using our 3D face reconstruction model fine-tuned with a two-stage strategy. One 3D face is created by estimating the 3DMM parameters, and the other 3D face is created by projecting the texture of the input image onto the estimated 3D shape. The rendered image $\mathcal{R}_c$ from the 3D face with the estimated texture is an occlusion-free facial image owing to the limited representation power of 3DMM and the rendered image $\mathcal{R}_p$ from the 3D face with the projected texture is a facial image that includes occlusion. Our key contribution is to provide strong self-supervision with a Swap-R&R strategy that extends the Rotate-and-Render strategy [55]. Specifically, the mask for the occlusion area is coarsely calculated based on the color and structural differences between the two face images. Then, the occlusion areas are exchanged between two rendered images by utilizing the calculated occlusion mask. Thus, $\mathcal{R}_c$ and $\mathcal{R}_p$ become the occluded and occlusion-free images, respectively. Next, we obtain a damaged facial image through two rotate-and-render operations of [55]. The process rotates a face in the 3D space back and forth, and re-renders it onto the 2D plane. $\mathcal{R}_p$ becomes a rendered facial image, with any random pose through a first rotate-and-render operation. A second rotate-and-render operation creates a facial image with the original pose. Finally, our generator learns to restore the original input image from two images, $\mathcal{R}_c$ and $\mathcal{R}_p$. The generator is designed to provide structural and textural information from $\mathcal{R}_c$ for the recovery of $\mathcal{R}_p$. In addition, we add an occlusion parsing path to focus on occluded and damaged regions so that more natural images can be recovered. On the other hand, in the inference process, our generator creates an image without occlusion from the rendered images of the two 3D faces.

In this paper, we propose a self-supervised joint face rotation and de-occlusion method that can recover a photorealistic occlusion-free facial image. Qualitative and quantitative results show that our method outperforms previous state-of-the-art methods for both constrained and in-the-wild images. In addition, our method does not require a paired training dataset. The contributions of this paper can be summarized as follows:

- We propose a novel face rotation and de-occlusion model that is guided by 3DMM and a coarse occlusion mask.
- We propose a novel Swap-R&R strategy for strong self-supervision that does not require paired training data for joint face rotation and de-occlusion.
- We present an occlusion-robust 3D face reconstruction model through two-stage fine-tuning.

2. Related Work

2.1. Face Rotation and Multi-View Synthesis

The face rotation task generates multi-view face images when given a single-view face image and required poses. Specifically, frontalization has received more attention, as it generates a frontal face image. Face rotation methods can be roughly divided into GAN-based methods and 3D-based methods.

GAN-based methods. DR-GAN [39] adopted a GAN to generate a frontal face with an encoder-decoder architecture for the first time. However, the generated results are unsatisfactory and contain serious artifacts. TP-GAN [20] utilizes global and local networks to consider individual facial components using a multi-task learning strategy. CAPG-GAN [18] leverages face heatmaps, which provide the location information of key facial components. Similar to TP-GAN, PIM [52] adopts a dual-path generator to produce high-quality images by adding regularization terms to effectively learn face representations. FNM [33] normalizes faces from the pose, expression, illumination and occlusion by training to combine labeled and unlabeled data, however, their results lead to overfitting of a constrained Multi-PIE dataset environment. Dual et al. [11] proposed a face de-occlusion and frontalization method using a boosting generator, but this method only performs face de-occlusion for white regions, not general objects. Furthermore, it fails to preserve their identities or generate high-quality results.

3D-based methods. FF-GAN [47] integrates a 3DMM coefficient regression network and generative model. The network acquires low- and high-frequency information from the 3DMM coefficients and an image, respectively. 3D-PIM [55] obtains prior information for pose, using 3DMM fitting and pose normalization. Then, a dual-path generator for facial components is used to generate a frontal face image. HF-PIM [5] generates high-quality frontalized facial images via facial texture maps and correspondence fields. However, a paired dataset is required for training. Rotate-and-Render [55] proposed a self-supervised framework for face rotation, which corrupts an image by applying two rotate-and-render operations and learns to reconstruct the image to the original image.
2.2. Face De-Occlusion and Completion

Image completion aims to fill the erased area of the image when given with the mask for the erased region. SC-FEGAN [21] achieves high-resolution in-painted face images by using color maps and sketches. Face de-occlusion automatically detects erased regions, as well as occlusion due to various factors and recovers the regions naturally.

Most face de-occlusion methods [51, 50, 4, 10] learn to reconstruct original images from synthesized images occluded with a limited set of objects. Zhao et al. [51] reconstructs de-occluded and identity-preserved face images using a CNN supervised with identity labels. Through an additional occlusion detection channel, an occlusion mask is calculated and combined with the reconstructed face. However, they only handle grayscale face images and generate results with artifacts. Yuan et al. [50] guides the facial structure with a 3DMM prior and uses local and global discriminators. However, the de-occluded area of the generated results is blurry, and this method fails to de-occlude face images with more than one type of occlusion. STN-GAN [45] fills the erased region around facial key components under the guidance of facial landmark points. As with Dual et al. [11], inpainting is performed only for constrained white areas.

3. Approach

Our self-supervised method for joint face rotation and de-occlusion comprises three parts: occlusion-robust 3D face reconstruction, Swap-R&R strategy, and generator for complete face recovery. Our model aims to recover a face corrupted by rotation and occlusion with the help of a 3D face.

3.1. Occlusion Robust 3D Face Reconstruction

The first step of our method is to regress the 3DMM coefficients from the input image. We use the original [9] as a baseline model for 3D face reconstruction and fine-tune the model in two stages, since existing 3D face reconstruction methods tend to show unnatural results in both shape and texture for occluded face images. We briefly summarize 3DMM and then introduce our novel fine-tuning strategy that makes the model robust to occlusion.

**3DMM.** In a 3DMM, the face shape \( S \) and the texture \( T \) can be represented as:

\[
S = S(\alpha, \beta) = \bar{S} + B_{id}\alpha + B_{exp}\beta,
\]

\[
T = T(\delta) = \bar{T} + B_{\delta}\delta,
\]

where \( \bar{S} \) and \( \bar{T} \) are the average face shape and texture; \( B_{id} \), \( B_{exp} \), and \( B_{\delta} \) are the PCA bases of identity, expression, and texture respectively, which are all scaled with standard deviations; \( \alpha \), \( \beta \), and \( \delta \) are the corresponding coefficient vectors for generating a 3D face. We adopt the 3DMM parameter regressor [9]. Given a face image, it regresses a 239-dimensional vector \( \{C, p, \gamma\} \). \( C \) consists of \( \alpha \in \mathbb{R}^{80}, \beta \in \mathbb{R}^{64}, \) and \( \delta \in \mathbb{R}^{80}, p \) is a 6-dimensional 3D face pose for rotation and translation, and \( \gamma \) is a 9-dimensional Spherical Harmonics (SH) [34]. The output 3D mesh contains 36K vertices excluding ear and neck areas.

**Occlusion-robust 3D face.** We propose our novel two-stage fine-tuning strategy for occlusion-robust 3D face reconstruction. The training method is split into two training stages due to the difficulty of initial training for extreme occlusions. We fine-tune the baseline with our newly created datasets in the first stage and with teacher-student learning method in the second stage.

For the first fine-tuning stage, we create two occluded face datasets. In order to train occlusion-robust 3D face model, occluded face image datasets are essential, but they are absent. So, we create datasets by synthesizing the hand-shaped mask on two datasets, 300W-LP [57] and CelebA [28]. The 300W-LP is synthesized dataset in extreme poses through 3D image rotation and CelebA is real face image dataset. The hand-shaped mask is randomly transformed with rotation and scaling and is located around the facial landmarks. When training, the model estimates the 3D face with the occluded face images as input, and all losses are calculated using the original image as the target. Furthermore, the Multi-PIE dataset [14] is used for robustness to various poses and illuminations. The landmarks for the Multi-PIE and CelebA datasets are estimated through 3DDFA-V2 [13]. We follow the overall loss function from [9], but we multiply 0.7 to facial landmark loss for CelebA and Multi-PIE dataset to prevent error propagation of 3DDFA-V2.

For the second fine-tuning stage, we introduce teacher-student training strategy on a Masked Face-Net dataset [3], since there still exists a limitation to the face image where extreme occlusion exists. The Masked Face-Net is a dataset created by synthesizing a dental mask on high-resolution face images in FFHQ dataset. Additionally, we use Random Erasing [54] on the FFHQ dataset, which randomly erases a few pixels in the image to avoid overfitting on the Masked-Face-Net dataset. Both the teacher model \( T \) and the student model \( S \) are initialized identically with the weights of the fine-tuned model in the first stage. \( T \) and \( S \) take images in the FFHQ and Masked Face-Net datasets as inputs, respectively. Therefore, the entire network is trained to predict occlusion-robust 3D faces. A combination of parameter loss, perceptual teaching loss, and landmark loss is used in the training process. To balance the terms, weights are set to 0.1, 1.0 and 0.01 in the order mentioned.

For parameter loss, we leverage coefficients regressed by the teacher network as the ground-truth.

\[
\mathcal{L}_{\text{para}}(T, S) = ||T(I) - S(I')||_2,
\]
where \( I' \) is the synthesized face image with a dental mask on input image \( I \).

Inspired by \cite{22}, we regularize the distance between features from the top \( K \) layers of the teacher and the student network. Our perceptual teaching loss is defined as

\[
\mathcal{L}_{\text{teach}}(T, S) = \frac{1}{K} \sum_{i=1}^{K} \omega_i \|T^{(i)}(I) - S^{(i)}(I')\|_2, \tag{3}
\]

where \( T^{(i)} \) and \( S^{(i)} \) represent the \( i \)th layer of each model. We leverage only the features from the top 4 layers prior to the fully connected layers. \( w_i \) is the weight of the feature distance of each layer, which we set to 0.125, 0.25, 0.5, and 1, respectively.

We also guide the 68 3D facial landmark locations. The 3D landmark vertices of the reconstructed 3D face are projected onto the 2D plane by function \( q \) and used to calculate the loss.

\[
\mathcal{L}_{\text{lan}}(x) = \frac{1}{N} \sum_{n=1}^{N} \omega_n \| q_n - q'_n(x) \|_1, \tag{4}
\]

where \( \omega_n \) is the landmark weight. We set the inner mouth and eye points to 20 and the others to 1. Through our novel fine-tuning strategy, the outputs are much more robust to the occlusion than the results from the baseline model.

### 3.2. Swap-R&R Strategy

We use our occlusion-robust 3D face reconstruction model to generate two different 3D faces. The first 3D face is generated from the shape and texture parameters estimated through our model. The second 3D face is created by projecting pixels of input image onto the estimated 3D shape. Then, the rendered image \( R_e \) from the first 3D face which has the estimated texture is an occlusion-free facial image owing to the limited representation power of 3DMM while the rendered image \( R_p \) from the second 3D face with projected texture includes occlusion.

Inspired by the Rotate-and-Render strategy \cite{55} for face rotation, we propose a swap-R&R strategy, which enables our model to be trained in a fully unsupervised manner for joint face frontalization and de-occlusion task. Our intuition is that a 3DMM-based reconstructed 3D face is an occlusion-free image and can guide the recovery for corrupted regions with large gaps in textual and structural features. First, we coarsely calculate the mask \( M \) for the occlusion area irrespective of the type of object by leveraging the structural and textural information. Face parsing networks cannot distinguish not-trained objects like hands. So, we only use it as an auxiliary role such as excluding the eye area. The occlusion mask \( M \) from the two rendered images can be acquired as follows:

\[
M = z_t z_s + z_t + \alpha z_s, \tag{5}
\]

where \( z_t \) and \( z_s \) are the z-scores of the texture differences \( d_t \) and structural differences \( d_s \) between two rendered images, respectively. \( \alpha \) is the weight to compromise between occlusion and skin details, such as wrinkles, and is empirically set to 0.4. Then, \( M \) is normalized with the mean and standard deviation, and areas with values above zero define as occlusion areas. We compute the textural differences \( d_t \) between \( R_p \) and \( R_e \) using L2 distance in the CIE-Lab color space \cite{36}. SSIM \cite{42} is used to calculate the structural differences, \( d_s \). We only use the product of contrast and structure, without calculating luminance. Finally, we add masks for eyeglasses and hairs and subtract eyes area using BiseNet \cite{48}. See the supplementary (Sec 2) for detailed formulations and descriptions.

Then, we swap the texture between \( R_p \) and \( R_e \) for the occlusion area that exists within \( R_p \) as follows:

\[
\begin{align*}
R_{ps} &= (1 - M) \otimes R_p + M \otimes R_e, \\
R_{es} &= (1 - M) \otimes R_e + M \otimes R_p,
\end{align*} \tag{6}
\]

after dilating and blurring \( M \) to synthesize naturally around the occlusion area. Additionally, to avoid referring to the texture of the rendered \( R_e \) as it is, we determine \( R_{es} \) via blurring for \( R_p \). Finally, \( R_{ps} \) is generated as a broken image through two Rotate-and-render operations. \( R_{ps} \) and \( R_{es} \) comprise a training pair for our overall network. Our swap-R&R strategy is illustrated as Fig 2.

### 3.3. CFR-GAN: Complete Face Recovery GAN

Our overall framework is illustrated in Fig 3. In the training stage, we take both rendered images \( R_{ps} \) and \( R_{es} \).
Figure 3: The overall framework of the proposed method. In the training stage, the network is trained to restore the input image from two images generated from Swap-R&R strategy. In the testing stage, the rotated and de-occluded face image is inferred from two rendered face images with any pose.

Figure 4: Generator architecture. $G$ includes generation path and occlusion parsing path. The occlusion parsing path consists of 3D-guided occlusion attention module and mixing module.

generated through the Swap-R&R framework as inputs and learn to reduce the differences between the generated image and original image $I$. This learning strategy allows the network to use $R_e$ to restore the collapsed textures and disappeared occlusion areas within $R_p$. In other words, the network is trained to consider the structurally and texturally different regions of $R_e$ and $R_p$ to be occlusion areas and restore those areas using the $R_e$. Simultaneously, our CFR-GAN is trained to recognize the location information for occlusions raised by inter-occlusion and self-occlusion. During the testing stage, thus, our model produces a high-quality rotated and de-occluded image and an occlusion mask by taking the rendered $R_p$ and $R_e$ as inputs, which are rotated only once to the desired pose. The details of the network are described below.

**Discriminator.** We employ a multi-scale discriminator from Pix2PixHD [41] and apply a gradient penalty $GP$ from WGAN-DIV [44] to stabilize the training of the GAN. Our loss function of discriminator $D$ and adversarial loss of generator $G$ are formulated as follows:

$$
\mathcal{L}_D = -\mathbb{E}(D(I)) + \mathbb{E}(D(G(R_p, R_e))) + GP,
$$

$$
\mathcal{L}_{adv} = -\mathbb{E}(D(G(R_p, R_e))).
$$

**Generator.** We employ CycleGAN [56], an image-to-image translation network, as the base structure of our generator $G$. The generator of CycleGAN is composed of a down-sampling module, residual blocks, and an upsampling module. We improve the down-sampling module to enable the detection and removal of occlusions by adding an occlusion parsing path $O$ and leveraging the spatial attention mechanism. The module enhances the feature representation of occlusions with the help of the $O$ path and enfeebles the corresponding features via the attention mechanism. A 3D-guided occlusion attention module and a mixing module in the $O$ path calculate the distance between $R_p$ and $R_e$ in the feature space and combine the gating feature maps extracted from the gated convolutions [49]. Specifically, the 3D-guided occlusion attention module is formulated as follows:

$$
f_o = IN((f_p - f_e)^2),
$$

$$
f_p = f_p \otimes (1 - \sigma(f_o)),
$$

where $f_p$ and $f_e$ are the input feature maps of $R_p$ and $R_e$, respectively, with the same spatial dimensions as input image $I$. $IN$ is instance normalization [40] and $f_p$ is newly updated using $f_o$. The details of our generator is depicted in Fig 4. To provide attention to occlusions and obtain an occlusion mask $M_o$, our model is optimized by targeting the coarse occlusion mask $M_o$, our model is optimized by targeting the coarse occlusion mask $M_o$ as the ground-truth. However, there may be problems in generating the facial components in the testing stage and inaccuracy of $M_o$ because the model is mostly learned to recover non-facial parts during the training stage. We alleviate the problems with simple data augmentation, RandomErasing. We obtain our model
to focus more on occlusion through the following loss functions: occlusion mask loss $L_m$ and occlusion-aware reconstruction loss $L_{rec}$.

$$L_m = \| M_o - M \|_2,$$  \hspace{1cm} (10)

$$L_{rec} = \frac{1}{N_M} \sum_{i=1}^{N_M} (M \odot \sum_{c=1}^{3} G(\mathcal{R}_{p_{cr}}, \mathcal{R}_{e_c}) - I)).$$  \hspace{1cm} (11)

$L_{rec}$ is defined as $L_1$ distances between the ground truth image $I$ and output image of the generator $G$, which is only calculated for the mask $M$. To regularize the distance between the output and target features, we use the perceptual loss $L_{per}$ using VGG-19 \[37\] network pre-trained from ImageNet. The loss function is calculated with the feature maps $F_{vgg}$ which are the outputs of $N_{vgg}$ layers in the VGG-19 network as follows:

$$L_{per} = \sum_{i=1}^{N_{vgg}} \| F_{vgg}^{(i)}(I) - F_{vgg}^{(i)}(G(\mathcal{R}_{p_{cr}}, \mathcal{R}_{e_c})) \|_1.$$  \hspace{1cm} (12)

When the network recovers the facial components in the corrupted regions, it tends to try to imitate $R_e$ as a guidance without the consideration for identities. So, to preserve identities, we add an identity loss function using a face recognition network. The face recognition network, which is ResNet-50 \[17\] is trained with ArcFace \[7\] on MS1M \[16\]. The loss function is:

$$L_{id} = 1 - \frac{F_{id}^I \cdot F_{id}^G}{\max(\|F_{id}^I\|_2 \cdot \|F_{id}^G\|_2, \epsilon)},$$  \hspace{1cm} (13)

where $F_{id}^I$ and $F_{id}^G$ are the 512-dimensional output vectors of the face recognition network for an input $I$ and the output image of $G$, respectively. $\epsilon$ sets to very small value 1e-8 to avoid division by zero.

Our total loss function $L$ of the generator is as follow:

$$L = L_{adv} + \lambda_{id} L_{id} + \lambda_{m} L_{m} + \lambda_{per} L_{per} + \lambda_{rec} L_{rec};$$  \hspace{1cm} (14)

where $\lambda$ is multiplied to balance the loss terms.

4. Experiments

4.1. Experimental Settings

Implementation Detail. For all input images, we perform face alignment based on the extracted eyes, nose, and mouth with \[6\]. Instance Normalization \[40\] and Spectral Normalization \[31\] are applied to all layers in $G$, except for the $O$ path. Discriminator $D$ is composed of two scales that use the same network structure with 6 CNN layers. For our base 3D face model, we only use R-Net without using C-Net from \[9\]. Each input image size for $R_{3D}$ and the generator is 224. The weights are updated using the AdamW optimizer \[29\]. The inference time is about 0.05s to generate a rotated and de-occluded face image from a single image when using 1 TITAN XP GPU. Pytorch3D \[35\] is also used as a 3D renderer for training and rendering. Please refer to our code for more information.

Datasets. Our approach does not depend on occlusion-free images and paired multi-view data because we provide strong self-supervision using the swap-R&R strategy. Therefore, we choose CelebA-HQ \[23\], CelebA \[28\], and FFHQ \[24\] for training the CFR-GAN, which are generally used for high-quality face datasets containing some occluded face images. To evaluate our boosting performance on face-related tasks, we test our methods on facial recognition and facial expression recognition. For the evaluation of facial recognition, LFW \[12\] and IJB-A \[2\] containing profile faces and occluded faces are used. Moreover, more difficult face recognition datasets such as IJB-B \[43\] and IJB-C \[50\] are used for additional comparison with the state-of-the-art model. To evaluate facial expression recognition, a large-scale facial expression database RAF-DB \[26\] is leveraged.

4.2. Qualitative Results

Results on challenging images. The results on face images with both extreme poses and complicated occlusions are illustrated in Fig. 5. We show that natural de-occluded face images with background can be obtained through combinations of original images and synthesized images with produced occlusion masks through $O$ path. Our model works well when complex or multiple types of occlusions exist simultaneously, too.

Comparison with face rotation methods. Fig. 6 illustrates the results of the face frontalization methods. Similar to our method, FF-GAN \[37\] and Rotate-and-Render combine 3D and GAN. However, FF-GAN fails to fully make the frontal face and has losses for regions of self-occlusion
Figure 6: Qualitative comparison with methods for face frontalization.

Figure 7: Qualitative comparisons with state-of-the-art methods of each task. The left side and the right side are results for face frontalization and de-occlusion, respectively.

Table 1: Verification performance (%) on LFW and IJB-A dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>LFW ACC / AUC(%)</th>
<th>IJB-A @FAR=.01 / .001</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP-GAN [20]</td>
<td>96.13 / 99.42</td>
<td>-</td>
</tr>
<tr>
<td>FF-GAN [47]</td>
<td>96.42 / 99.45</td>
<td>85.2 / 66.3</td>
</tr>
<tr>
<td>DR-GAN [39]</td>
<td>-</td>
<td>87.2 / 78.1</td>
</tr>
<tr>
<td>CAPG-GAN [18]</td>
<td>99.37 / 99.90</td>
<td>-</td>
</tr>
<tr>
<td>FNM [33]</td>
<td>-</td>
<td>93.4 / 83.8</td>
</tr>
<tr>
<td>Res18 [17]</td>
<td>98.85 / 99.90</td>
<td>90.57 / 80.0</td>
</tr>
<tr>
<td>Res18+R&amp;R [55]</td>
<td>98.95 / 99.91</td>
<td>91.98 / 82.48</td>
</tr>
<tr>
<td>Res18+Ours</td>
<td>99.23 / 99.92</td>
<td>93.36 / 82.88</td>
</tr>
</tbody>
</table>

Table 2: 1:1 Verification performance (TAR@FAR) on the IJB-B and IJB-C dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>IJB-B @FAR=.01 / .001</th>
<th>IJB-C @FAR=.01 / .001</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASIA+Ours</td>
<td>83.17 / 47.42</td>
<td>81.41 / 37.09</td>
</tr>
<tr>
<td>CASIA R&amp;R [55]</td>
<td>82.68 / 21.60</td>
<td>79.39 / 18.80</td>
</tr>
<tr>
<td>CASIA Ours w/o $L_{id}$</td>
<td>71.30 / 0</td>
<td>36.55 / 6.48</td>
</tr>
<tr>
<td>CASIA Ours</td>
<td>81.90 / 67.08</td>
<td>83.46 / 68.89</td>
</tr>
<tr>
<td>CASIA Ours</td>
<td>85.34 / 73.54</td>
<td>86.46 / 74.81</td>
</tr>
</tbody>
</table>

Table 3: Test accuracy (%) on RAF-DB dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Total Occlusion Pose(&gt; 30)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG16 [37]</td>
<td>82.53 / 76.87 / 79.39</td>
</tr>
<tr>
<td>GACNN [27]</td>
<td>85.07 / 80.54 / -</td>
</tr>
<tr>
<td>VGG16+ours</td>
<td>85.48 / 80.54 / 84.05</td>
</tr>
</tbody>
</table>

Table 1: Verification performance (%) on LFW and IJB-A dataset.

Table 2: 1:1 Verification performance (TAR@FAR) on the IJB-B and IJB-C dataset.

Table 3: Test accuracy (%) on RAF-DB dataset.

such as TP-GAN [20]. HF-PIM and Rotate-and-Render produce high-quality results similar to ours but some occlusions, such as hairs and a finger, remain in their results. CAPG-GAN [19] and FNM [33] generate results in which the identity is not preserved. The results of TP-GAN [20], DR-GAN [39], and Res18 [17] in Fig. 6 are extracted from [55]. Compared to these studies, our results seem to be clearer. Through the additional results in Fig. 7, our method exhibits better performance than Rotate-and-Render for occlusion-free face images.

Comparison with face de-occlusion methods. Most studies on face de-occlusion do not offer their codes and models. However, our results can be compared with published results from [55], as illustrated in Fig. 7. Our results show structurally more complete results for the de-occluded part than their results. To further clarify our contributions, we show the results for images with more than one type of occlusion mentioned as a limitation in [50], as well as the results for the case where complex occlusions exist, as illustrated in Fig. 5.

Because it is nearly impossible to find a perfectly frontal and occlusion-free face image to use as the ground truth, it is difficult to measure accurate numerical results. Furthermore, we only limit the range of occluded regions to the face area. Therefore, we substitute the numerical results with the results of additional experiments for facial recognition and facial expression recognition.

4.3. Extensive Experiments

Facial Recognition To evaluate our method on facial recognition task, we compare the verification performance with frontalization methods on LFW and IJB-A datasets. However, as mentioned on [55], previous methods do not follow either clear setting or using different baseline (e.g. LightCNN29) which is not comparable. Therefore, we follow settings of [55], which uses ResNet18 [17] as backbone, ArcFace [21] for loss function, and CASIA-WebFace for training data. First, we evaluate the performance on LFW and IJB-A datasets with the recognition model, which is trained with the dataset augmented via our method like the previous method. The results are listed in table 1. Moreover, we demonstrate that our model remarkably boosts the performance on harder dataset, IJB-B and IJB-C, which are mostly used in the facial recognition task. We also expand our experiments by adding the performance when preprocessing testing datasets with our method. As the results listed in table 2, we boost the facial recognition perfor-
Table 4: The NME (%) on AFLW2000-3D dataset (68 pts).
The rows of “Ori” and “Occ” present NME for original images and occluded facial images, respectively.

Figure 8: Ablation study on occlusion-robust 3D face reconstruction. The 3D faces estimated by [9] and ours are shown on second and third columns, respectively.

Figure 9: Ablation study on our overall method. Ours w/o \{diff\} is results only using face parsing network, not our entire algorithm to calculate a coarse occlusion mask $M$.

4.4. Ablation Study

Occlusion-robust 3D face To validate the effectiveness of occlusion-robust face reconstruction, we present the results according to fine-tuning in Fig. 8. The results show that fine-tuned model can better estimate the shape and texture of faces existing severe occlusions. For quantitative results, Normalized Mean Error (NME) is measured for AFLW2000-3D database to evaluate 3D face alignment, as shown in table 4. We evaluate the robustness for occlusion by evaluating for both the original images and the occluded face images synthesized with hand-shaped masks.

CFR-GAN Fig. 9 shows that our methods strongly affect removing occlusions. When the training data was generated without swap in swap-R&R, most occlusions except to occlusions by rotation remain remarkably. The results for the model trained without $O$ path were only erased for objects with a strong difference. To verify help detect diverse occlusions our algorithm to calculate a coarse occlusion mask, the mask is calculated by only using a face parsing network. The results are better than previous cases, but occlusions like hands not classified by the face parsing network were not completely removed. Additionally, through comparison with Rotate-and-render [55] which is a face rotation method, we show the necessity of joint face rotation and de-occlusion. It is difficult to apply an additional face de-occlusion method because of the remaining afterimages of occlusions on frontalized results and the collapse of facial structure. In experiments for the absence of $O$ path, we changed the occlusion-aware reconstruction loss to a reconstruction loss for a total image due to not generate an occlusion mask.

5. Conclusion

In this paper, we present a CFR-GAN for joint face rotation and de-occlusion. Unlike existing methods, which suffer from the lack of high-quality datasets, our method does not require paired dataset. We provide a strong self-supervision by synthesizing a damaged face image with our occlusion-robust 3D reconstruction model and Swap-R&R strategy. Our method outperforms previous state-of-the-art methods for qualitative results. Furthermore, this work can boost the performance for other face-related tasks and be a step forward regarding training joint face rotation and de-occlusion networks in a fully unsupervised manner.

6. Acknowledgement

This work was supported by Institute of Information & communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program(Korea University))
References


[27] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with

