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

Yeong-Joon Ju[1*] Gun-Hee Lee[2*] Jung-Ho Hong[1] Seong-Whan Lee[1†]
[1]Department of Artificial Intelligence, Korea University, Seoul, South Korea
[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[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-occlusion can alleviate these problems but are challenging tasks because of the lack of high-quality training data.

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.

Most traditional methods for face rotation use the 3D Morphable Model (3DMM) [1], a statistical model of facial shape and texture that uses a set of linear basis functions [8, 46, 38]. A challenging issue is the natural estimation of the texture of the invisible face area. Zhu *et al.* [46] 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 [53, 47, 55] have been proposed to synthesize photorealistic rotated faces by utilizing the power of the Generative Adversarial Network (GAN) [13]. 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 [51, 11, 50, 4, 10] 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

*Equal contribution. †Corresponding author.
[1] https://github.com/yeongjoonJu/CFR-GAN.

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 $\mathcal{R}_e$ 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, $\mathcal{R}_e$ 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, $\mathcal{R}_e$ and $\mathcal{R}_p$. The generator is designed to provide structural and textural information from $\mathcal{R}_e$ 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 [53] 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:

$$\begin{aligned} S &= S(\alpha, \beta) = \bar{S} + B_{id}\alpha + B_{exp}\beta, \\ T &= T(\delta) = \bar{T} + B_t\delta, \end{aligned} \tag{1}$$

where $\bar{S}$ and $\bar{T}$ are the average face shape and texture; B_{id}, B_{exp}, and B_t are the PCA bases of identity, expression, and texture respectively, which are all scaled with standard deviations; α, β, and δ 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 γ 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-shape 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 [15]. 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-FaceNet dataset. Both the teacher model $\mathcal{T}$ and the student model $\mathcal{S}$ are initialized identically with the weights of the fine-tuned model in the first stage. $\mathcal{T}$ and $\mathcal{S}$ 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.

$$\mathcal{L}_{para}(\mathcal{T}, \mathcal{S}) = \|\mathcal{T}(I) - \mathcal{S}(I')\|_2, \tag{2}$$

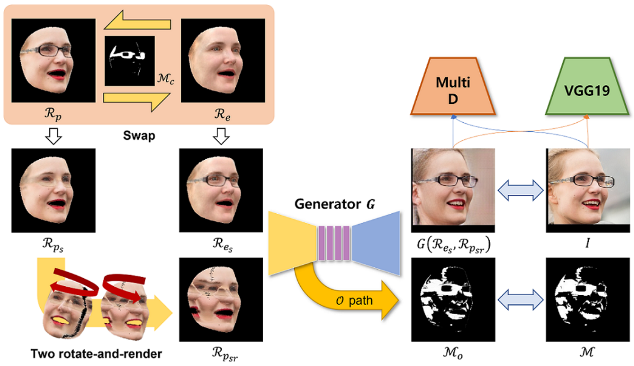

Figure 2: **Swap-R&R strategy to generate training pairs for self-supervision.** We **swap** $\mathcal{R}_p$ and $\mathcal{R}_e$ for the corresponding regions on a coarse occlusion mask $\mathcal{M}$. Then, $\mathcal{R}_{p_{sp}}$ is generated from the swapped $\mathcal{R}_{p_s}$ via two rotate-and-render operations.

where I' is the synthesized face image with a dental mask on input image I.

Inspired by [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

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

where $\mathcal{T}^{(i)}$ and $\mathcal{S}^{(i)}$ represent the ith 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.

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

where ω_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 $\mathcal{R}_e$ 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 $\mathcal{R}_p$ from the second 3D face with projected texture includes occlusion.

Inspired by the Rotate-and-Render strategy [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 $\mathcal{M}$ 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 $\mathcal{M}$ from the two rendered images can be acquired as follows:

$$\mathcal{M} = z_t z_s + z_t + \alpha z_s, \quad (5)$$

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. α is the weight to compromise between occlusion and skin details, such as wrinkles, and is empirically set to 0.4. Then, $\mathcal{M}$ 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 $\mathcal{R}_p$ and $\mathcal{R}_e$ using L2 distance in the CIE-Lab color space [36]. SSIM [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 [48]. See the supplementary (Sec 2) for detailed formulations and descriptions.

Then, we swap the texture between $\mathcal{R}_p$ and $\mathcal{R}_e$ for the occlusion area that exists within $\mathcal{R}_p$ as follows:

$$\begin{aligned} \mathcal{R}_{p_s} &= (1 - \mathcal{M}) \otimes \mathcal{R}_p + \mathcal{M} \otimes \mathcal{R}_e, \\ \mathcal{R}_{e_s} &= (1 - \mathcal{M}) \otimes \mathcal{R}_e + \mathcal{M} \otimes \mathcal{R}_p, \end{aligned} \quad (6)$$

after dilating and blurring $\mathcal{M}$ to synthesize naturally around the occlusion area. Additionally, to avoid referring to the texture of the rendered $\mathcal{R}_{e_s}$ as it is, we determine $\mathcal{R}_{e_s}$ via blurring for $\mathcal{R}_p$. Finally, $\mathcal{R}_{p_{sr}}$ is generated as a broken image through two Rotate-and-render operations. $\mathcal{R}_{p_{sr}}$ and $\mathcal{R}_{e_s}$ 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 $\mathcal{R}_{p_{sr}}$ and $\mathcal{R}_{e_s}$

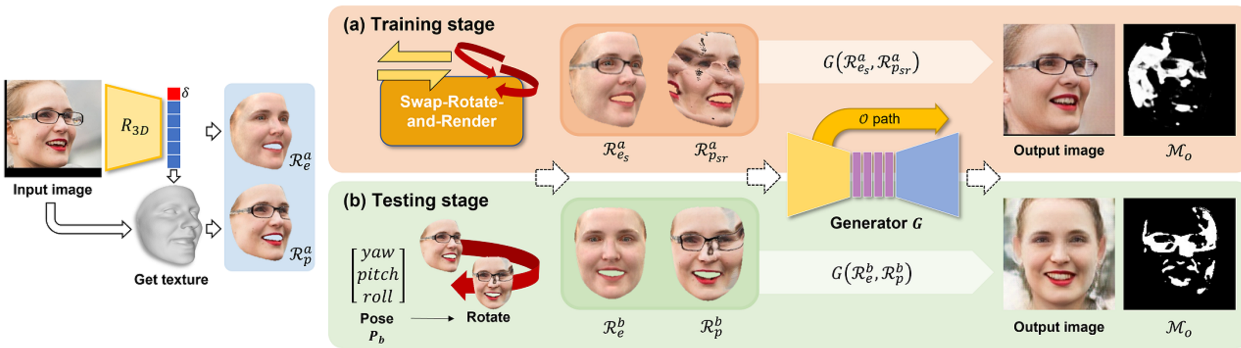

Figure 3: **The overall framework of the proposed method.** In the training stage, the network is trained to restore the input image from two images generated from Swap-R&R strategy. In the testing stage, the rotated and de-occluded face image is inferred from two rendered face images with any pose.

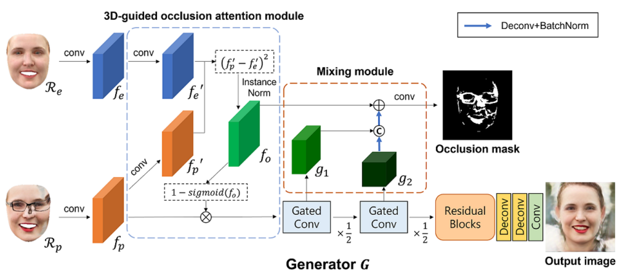

Figure 4: **Generator architecture.** G includes generation path and occlusion parsing path. The occlusion parsing path consists of 3D-guided occlusion attention module and mixing module.

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 $\mathcal{R}_e$ to restore the collapsed textures and disappeared occlusion areas within $\mathcal{R}_p$. In other words, the network is trained to consider the structurally and texturally different regions of $\mathcal{R}_e$ and $\mathcal{R}_p$ to be occlusion areas and restore those areas using the $\mathcal{R}_e$. 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 $\mathcal{R}_{p_s}$ and $\mathcal{R}_{e_s}$ 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:

$$\mathcal{L}_D = -\mathbb{E}(D(I)) + \mathbb{E}(D(G(\mathcal{R}_{p_{sr}}, \mathcal{R}_{e_s}))) + GP, \tag{7}$$

$$\mathcal{L}_{adv} = -\mathbb{E}(D(G(\mathcal{R}_{p_{sr}}, \mathcal{R}_{e_s}))). \tag{8}$$

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 up-sampling module. We improve the down-sampling module to enable the detection and removal of occlusions by adding an occlusion parsing path $\mathcal{O}$ and leveraging the spatial attention mechanism. The module enhances the feature representation of occlusions with the help of the $\mathcal{O}$ path and enfeebles the corresponding features via the attention mechanism. A 3D-guided occlusion attention module and a mixing module in the $\mathcal{O}$ path calculate the distance between $\mathcal{R}_p$ and $\mathcal{R}_e$ 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:

$$\begin{aligned} f_o &= IN((f_p - f_e)^2), \\ f_p &= f_p \otimes (1 - \sigma(f_o)), \end{aligned} \tag{9}$$

where f_p and f_e are the input feature maps of $\mathcal{R}_p$ and $\mathcal{R}_e$, 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 $\mathcal{M}_o$, our model is optimized by targeting the coarse occlusion mask $\mathcal{M}$ as the ground-truth. However, there may be problems in generating the facial components in the testing stage and inaccuracy of $\mathcal{M}$ 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 $\mathcal{L}_m$ and occlusion-aware reconstruction loss $\mathcal{L}_{rec}$.

$$\mathcal{L}_m = \|\mathcal{M}_o - \mathcal{M}\|_2, \tag{10}$$

$$\mathcal{L}_{rec} = \frac{1}{N_{\mathcal{M}}} \sum_{i=1}^{N_{\mathcal{M}}} (\mathcal{M} \odot \sum_{c=1}^{3} |G(\mathcal{R}_{p_{sr}}, \mathcal{R}_{e_s}) - I|). \tag{11}$$

$\mathcal{L}_{rec}$ 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 $\mathcal{M}$. To regularize the distance between the output and target features, we use the perceptual loss $\mathcal{L}_{per}$ 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:

$$\mathcal{L}_{per} = \sum_{i=1}^{N_{vgg}} \|F_{vgg}^{(i)}(I) - F_{vgg}^{(i)}(G(\mathcal{R}_{p_{sr}}, \mathcal{R}_{e_s}))\|_1. \tag{12}$$

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:

$$\mathcal{L}_{id} = 1 - \frac{F_{id}^I \cdot F_{id}^G}{max(\|F_{id}^I\|_2 \cdot \|F_{id}^G\|_2, \epsilon)}, \tag{13}$$

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. ϵ sets to very small value 1e-8 to avoid division by zero.

Our total loss function $\mathcal{L}$ of the generator is as follow:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_{id}\mathcal{L}_{id} + \lambda_m\mathcal{L}_m + \lambda_{per}\mathcal{L}_{per} + \lambda_{rec}\mathcal{L}_{rec}, \tag{14}$$

where λ 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 $\mathcal{O}$ 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

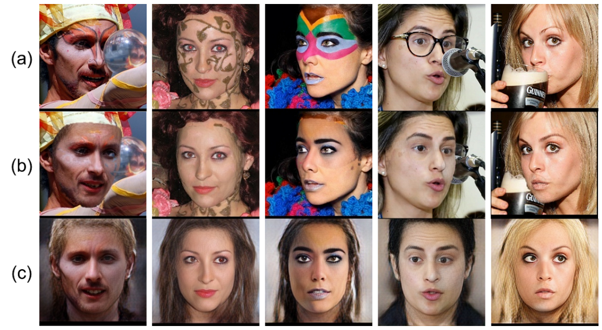

Figure 5: Our joint frontalization and de-occlusion results. (a) Input images. (b) De-occluded images. (c) Joint rotated and de-occluded results.

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. Pytorch [32] is used for the code implementation. 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 [30] 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, 25] 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 $\mathcal{O}$ 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 [47] 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.

Method	LFW ACC / AUC(%)	IJB-A @FAR=.01 / .001
TP-GAN [20]	96.13 / 99.42	-
FF-GAN [47]	96.42 / 99.45	85.2 / 66.3
DR-GAN [39]	-	87.2 / 78.1
CAPG-GAN [18]	99.37 / 99.90	-
FNM [33]	-	93.4 / 83.8
HF-PIM [5]	99.41 / 99.92	95.2 / 89.7
Res18 [17]	98.85 / 99.90	90.57 / 80.0
Res18+R&R [55]	98.95 / 99.91	91.98 / 82.48
Res18+Ours	**99.23 / 99.92**	**93.36 / 82.88**

Table 1: Verification performance (%) on LFW and IJB-A 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 [18, 5, 33] 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 [50], as illustrated in Fig. 7. Our results show

Train Set	Method	IJB-B @FAR=.01 / .001	IJB-C @FAR=.01 / .001
CASIA+Ours	-	83.17 / 47.42	81.41 / 37.09
CASIA	-	82.68 / 21.60	79.39 / 18.80
	R&R [55]	71.30 / 0	36.55 / 6.48
	Ours w/o $\mathcal{L}_{id}$	81.90 / 67.08	83.46 / 68.89
	Ours	**85.34 / 73.54**	**86.46 / 74.81**

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

Method	Total	Occlusion	Pose($>$ 30)
VGG16 [37]	82.53	76.87	79.39
GACNN [27]	85.07	80.54	-
VGG16+ours	**85.48**	80.54	**84.05**

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

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] for backbone, ArcFace [7] 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-

Method		[0,30]	[30,60]	[60,90]	Mean / Std
[9]	Ori	2.808	**3.332**	4.318	3.486 / 0.626
	Occ	3.949	5.530	7.335	5.605 / 1.383
Ours	Ori	**2.662**	3.404	**4.090**	**3.385** / 0.583
	Occ	**2.773**	**3.579**	**4.362**	**3.571** / 0.649

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.

mance more than the state-of-the-art model, giving margins of the boosting clearly. In addition, it can be mentioned that $\mathcal{L}_{id}$ prevents imitating R_e without the consideration for identities and helps preserve identities.

Facial Expression Recognition Table 3 shows the performance of the models on facial expression recognition task. We use the VGG-16 as a baseline model and add our model to identify performance change accordingly. GACNN [27] proposed additional modules to consider the occlusion by encoding patches from VGG-16. When our method is used for data augmentation, test accuracy remarkably increased, which is comparable to GACNN [27]. This result verifies that our joint face rotation and de-occlusion method can alleviate the problems of extreme pose and occlusion, and boost the performance on facial expression recognition task.

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 $\mathcal{O}$ 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

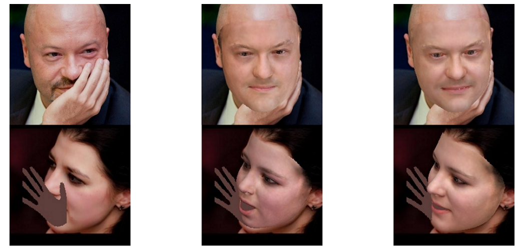

Figure 8: Ablation study on occlusion-robust 3D face reconstruction. The 3D faces estimated by [9] and ours are shown on second and third columns, respectively.

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 $\mathcal{M}$.

of occlusions on frontalized results and the collapse of facial structure. In experiments for the absence of $\mathcal{O}$ 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

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[2] Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015.

[3] Adnane Cabani, Karim Hammoudi, Halim Benhabiles, and Mahmoud Melkemi. Maskedface-net – a dataset of correctly/incorrectly masked face images in the context of covid-19. *Smart Health*, 2020.

[4] Jiancheng Cai, Hu Han, Jiyun Cui, Jie Chen, Li Liu, and S Kevin Zhou. Semi-supervised natural face de-occlusion. *IEEE Transactions on Information Forensics and Security*, 16:1044–1057, 2020.

[5] Jie Cao, Yibo Hu, Hongwen Zhang, Ran He, and Zhenan Sun. Learning a high fidelity pose invariant model for high-resolution face frontalization. *arXiv:1806.08472*, 2018.

[6] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

[8] Weihong Deng, Jiani Hu, Zhongjun Wu, and Jun Guo. Lighting-aware face frontalization for unconstrained face recognition. *Pattern Recognition*, 68:260–271, 2017.

[9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[10] Jiayuan Dong, Liyan Zhang, Hanwang Zhang, and Weichen Liu. Occlusion-aware gan for face de-occlusion in the wild. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020.

[11] Qingyan Duan and Lei Zhang. Look more into occlusion: realistic face frontalization and recognition with boostgan. *IEEE transactions on neural networks and learning systems*, 2020.

[12] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49 University of Massachusetts, 2007.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[14] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010.

[15] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[16] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[18] Yibo Hu, Xiang Wu, Bing Yu, Ran He, and Zhenan Sun. Pose-guided photorealistic face rotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 8398–8406, 2018.

[19] Yibo Hu, Xiang Wu, Bing Yu, Ran He, and Zhenan Sun. Pose-guided photorealistic face rotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 2439–2448, 2017.

[21] Youngjoo Jo and Jongyoul Park. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision (ECCV)*, pages 694–711, 2016.

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv:1710.10196*, 2017.

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[25] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019.

[26] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593, 2017.

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

attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018.

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017.

[30] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018.

[31] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv:1802.05957*, 2018.

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035. 2019.

[33] Yichen Qian, Weihong Deng, and Jiani Hu. Unsupervised face normalization with extreme pose and expression in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9851–9858, 2019.

[34] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500, 2001.

[35] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

[36] Alan R Robertson. The cie 1976 color-difference formulae. *Color Research & Application*, 2(1):7–11, 1977.

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[38] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015.

[39] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1415–1424, 2017.

[40] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv:1607.08022*, 2016.

[41] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[43] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus benchmark-b face dataset. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 90–98, 2017.

[44] Jiqing Wu, Zhiwu Huang, Janine Thoma, Dinesh Acharya, and Luc Van Gool. Wasserstein divergence for gans. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[45] Yifan Wu, Vivek Singh, and Ankur Kapoor. From image to video face inpainting: Spatial-temporal nested gan (stn-gan) for usability recovery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2396–2405, 2020.

[46] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 787–796, 2015.

[47] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 3990–3999, 2017.

[48] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[49] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4471–4480, 2019.

[50] Xiaowei Yuan and In Kyu Park. Face de-occlusion using 3d morphable model and generative adversarial network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10062–10071, 2019.

[51] Fang Zhao, Jiashi Feng, Jian Zhao, Wenhan Yang, and Shuicheng Yan. Robust lstm-autoencoders for face de-occlusion in the wild. *IEEE Transactions on Image Processing*, 27(2):778–790, 2017.

[52] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, and et al. Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2207–2216, 2018.

[53] Jian Zhao, Lin Xiong, Yu Cheng, Yi Cheng, Jianshu Li, Li Zhou, Yan Xu, Jayashree Karlekar, Sugiri Pranata, Shengmei Shen, and et al. 3d-aided deep pose-invariant face recognition. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2018.

[54] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 13001–13008, 2020.

[55] Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5911–5920, 2020.

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 2223–2232, 2017.

[57] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.