HAL open science

Segmentation-Reconstruction-Guided Facial Image De-occlusion

Xiangnan Yin, Di Huang, Zehua Fu, Yunhong Wang, Liming Chen

To cite this version:

Xiangnan Yin, Di Huang, Zehua Fu, Yunhong Wang, Liming Chen. Segmentation-Reconstruction-Guided Facial Image De-occlusion. 2022. hal-03482857

HAL Id: hal-03482857

https://hal.science/hal-03482857v1

Preprint submitted on 11 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:2112.08022v1 [cs.CV] 15 Dec 2021

Segmentation-Reconstruction-Guided Facial Image De-occlusion

Xiangnan Yin[1], Di Huang[2], Zehua Fu[2], Yunhong Wang[2], Liming Chen[1]
[1] Ecole Centrale de Lyon [2] Beihang University
{yin.xiangnan,liming.chen}@ec-lyon.fr, {dhuang,yhwang}@buaa.edu.cn, zehua_fu@163.com

Abstract

Occlusions are very common in face images in the wild, leading to the degraded performance of face-related tasks. Although much effort has been devoted to removing occlusions from face images, the varying shapes and textures of occlusions still challenge the robustness of current methods. As a result, current methods either rely on manual occlusion masks or only apply to specific occlusions. This paper proposes a novel face de-occlusion model based on face segmentation and 3D face reconstruction, which automatically removes all kinds of face occlusions with even blurred boundaries,e.g., hairs. The proposed model consists of a 3D face reconstruction module, a face segmentation module, and an image generation module. With the face prior and the occlusion mask predicted by the first two, respectively, the image generation module can faithfully recover the missing facial textures. To supervise the training, we further build a large occlusion dataset, with both manually labeled and synthetic occlusions. Qualitative and quantitative results demonstrate the effectiveness and robustness of the proposed method.

1. Introduction

Thanks to deep Convolutional Neural Networks (CNNs), massive training data, and the widespread use of cameras, face-related techniques have successfully found many applications in our daily life, covering the fields of entertainment, media, art, security, etc. However, face images in the wild may be occluded by various objects, which leads to loss of information and undesirable noise, further resulting in degraded performance of algorithms for face analysis. To address this issue, researchers typically combine methods such as augmenting training data with synthesized occlusions [1, 2], designing more sophisticated metrics and network structures [3–6], exploiting elaborated training strategies [7–9]. Despite their effectiveness, most of them are task-specific and cannot be migrated to other face-related tasks. A relatively more general solution is to de-occlude the face image before passing it to the downstream tasks.

Figure 1. Results of the proposed method. The first row and the third row are input images, the second row and the last row show the de-occlusion results. Zoom in for a better view.

Before the widespread use of deep learning-based methods, the dominant approaches were based on matching and copying visible patches to the missing parts. A representative one is PatchMatch [10], which recursively searches the nearest neighbor textures to fill in the holes. Such copy-and-past-based methods work well in recovering simple low-frequency textures and obtain smooth results; however, they cannot recover high-level textures with complex structures, e.g., eyes, nose, mouth. [11–14] leveraged Principal Component Analysis (PCA) to recover high-level semantic face textures. Nevertheless, such methods are based on the assumption that face images are in a linear space. Even for images acquired under constrained conditions, the results are blurry and far from satisfactory. As pointed out by [15], the deep autoencoder is a non-linear generalization of the PCA. Therefore, it is reasonable to replace the PCA part of the above methods with an autoencoder. SSDA [16] proposed a deep sparse autoencoder to remove Gaussian noises and superimposed text from images, pioneered the use of deep neural networks in this area. Influenced by the great success of Generative Adversarial Networks (GAN) [17] in image generation, Context Encoder [18] introduced adversarial training to image inpainting for the first time. Its im-

pressive performance led to the prosperity of GAN-based image inpainting methods. Since then, the combination of autoencoder and discriminator has been adopted as the basic model structure for image inpainting tasks [19–22]. Despite the remarkable progress so far achieved, current methods mainly focus on filling the holes with visually plausible textures and neglect the detection of imperfectness. When applied to the face de-occlusion task, images must be accompanied by manually labeled masks. One strategy to address this problem is to train the model to detect the occlusions automatically. Due to the lack of well-labeled paired training data, existing methods mainly train on the synthetic occluded face images [23–28]. Unfortunately, these methods either generate low-resolution results or only tackle specific occlusion types. We attribute this to the limited variety of synthetic occlusions. Real occlusions have different shapes, textures, and blurred boundaries, making it infeasible for synthesis-based methods to cover all possibilities of occlusion. In addition, collecting completely occlusion-free face images as ground truth is not easy because the forehead part of the face is often occluded by bangs, which is perhaps why none of the existing methods consider hair as a source of occlusion.

Given the limitations of existing methods, this paper explores how to enable neural networks to remove arbitrary kinds of face occlusions. Our main contributions are summarized as follows:

- We propose a novel face de-occlusion framework robust to all types of occlusions, regardless of their shapes and textures.
- We build a large occlusion dataset with extensive manually labeled occlusions from real face images.
- The proposed method outperforms the state-of-the-art baselines quantitatively and qualitatively.

2. Related Work

2.1. Mask-Dependent Face Inpainting

Image inpainting aims to recover the missing textures of the input image. Recently, the autoencoder-GAN-based models have achieved impressive results. Benefiting from such structure, Context Encoder [18] successfully predicts the 64×64 missing part of 128×128 images. To make the model further concentrate on the missing parts, [19, 20] leverage two discriminators: a global discriminator, which takes the whole image as input, and a local discriminator, which only takes the small region around the missing part. However, such a design is only suitable for a single rectangular(or even square) hole and cannot be applied to images with irregular and arbitrarily distributed holes. To properly handle free-form masks, [21] proposed a Partial Convolutional Layer, comprising a masked and re-normalized convolution operation followed by a mask-update step. [22] proposed a Gated Convolution module, which learns a soft mask for the features in different layers. However, due to the ineffectiveness of CNNs in modeling long-range correlations between distant textures and the hole regions, the inpainting results often have boundary artifacts and look unreal. [20] proposed a Contextual Attention module, which utilizes the image features as a convolution kernel, forcing the distant textures to interact with each other; this could be simply achieved by attention. [29] proposed a Self Attention module, which allows long-range dependency modeling of features; since then the attention mechanism has been widely applied in image generation tasks [30–33]. To make the generated image faithfully recover the facial topology, [34] leverages facial landmarks as a shape prior. [33] proposed a two-path probabilistic framework to generate pluralistic inpainting results. The common problem of the above methods is that they cannot detect the occluded parts automatically and rely on manually specified masks, limiting their usage scenarios.

2.2. Mask Free Face De-occlusion

Some methods are devoted to getting rid of masks' dependence and making the face de-occlusion process fully automated. They are typically trained on face pairs with synthesized occlusions. [23] proposed an LSTM-Autoencoder to gradually substitute the occlusions with facial textures. Benefiting from an elaborated forward pass, [24] proposed a semi-supervised face de-occlusion method, which is trained to predict the occlusion mask without the supervision of ground-truth. [25] proposed a two-stage GAN, where the first GAN aims to reconstruct the occlusion part solely, and the second GAN takes the result of the first GAN as a hint making the input image occlusion-free. Due to the limited variety of occlusions in the training set, the above methods can only generate images in 128×128 resolution with noticeable artifacts, restricting their application to face recognition. Given the complexity of occlusion types, a relatively simple task would be to deal with only specific occlusion: [27, 28] for eyeglasses and [35] for masks. Besides being limited by the occlusion variety, as they lack face topology information, the results are often far from satisfactory, especially on large-angled face images.

2.3. 3D-Guided Face De-occlusion

The work closest to ours is FD3D [26], where we both use 3D Morphable Model [36] (3DMM)-based reconstruction results as guidance. FD3D feeds the synthetic occluded face and the 3D reconstructed one to an adversarial auto-encoder and expects the output to be occlusion-free. Our method surpasses theirs in the following aspects: 1) They inpaint directly on the occluded face despite the chal-

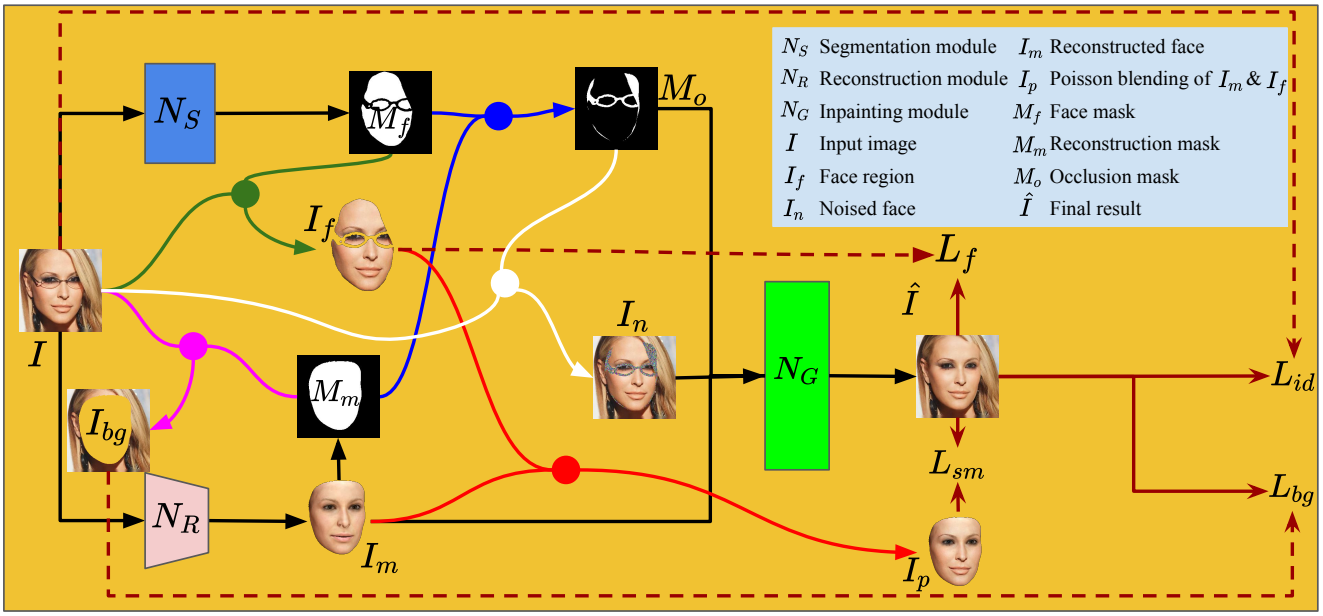

Figure 2. Overview of the proposed method, which does not include the discriminator part for space limitation

lenge of converting various occlusion textures to face textures; in contrast, our method explicitly removes occlusions with predicted masks before inpainting, bypassing the disturbance of the occlusion textures. 2) FD3D entirely relies on supervised learning, which requires occlusion-free images to serve as the ground truth, while our method can be trained on initially occluded images. 3) They synthesize images with fixed occlusions, namely, cup, glasses, hand, mask, and scarf, which have 99 variations in total; by detecting the face mask instead of innumerable types of occlusions, our method can handle all types of occlusion.

3. Proposed Method

3.1. Overview

Figure 2 illustrates the overview of our method. Given that covering all possible occlusions is impossible, we consider the opposite direction. We employ a face segmentation module N_S to predict the face mask M_f, which is much easier than predicting the mask of all kinds of occlusions. Then, we use the 3D reconstruction module N_R to predict the prior 3DMM-based texture and the corresponding mask for the whole face, denoted as I_m and M_m, respectively. The mask of the occlusions M_o can then be calculated as:

$$M_o = M_m - M_m \odot M_f, \tag{1}$$

where $\odot$ denotes element-wise product. For face inpainting, we replace the occluded part of the input image I with Gaussian noise to obtain I_n, which minimizes the disturbance of the occlusion textures. I_n, I_m, and M_o are concatenated and fed into the inpainting module N_G, and the output $\hat{I}$ is expected to be occlusion-free. The following terms supervise the inpainting: 1) the background region I_{bg}, 2) the non-occluded face region I_f, 3) the Poisson blending result of I_f and I_m, denoted as I_p.

The overall framework contains three modules, N_S, N_R, and N_G, for face segmentation, 3D reconstruction, and image inpainting. We present the implementation details of each module separately in the following.

3.2. Face Segmentation

Extracting face regions from occluded face images is a simple, valuable, but never seriously addressed problem. As a result, the recent 3D face reconstruction method [37] requires this functionality still needs to resort to a Naive Bayesian classifier. We present our solution to the problem in this section.

Data Preparation. We build our training data based on CelebAMask-HQ [38], which has 30,000 high-resolution face images, each with a segmentation mask of facial attributes. Basically, given a segmentation mask of a face image, we can easily get the face region by gathering the labels of eyes, nose, lips, mouth, and skin. Unfortunately, the segmentation mask in the dataset does not consider the occlusions, e.g., hands, accessories, microphones, which are all misclassified as skin. In addition, the segmentation masks in the dataset do not distinguish between sunglasses and spectacles; however, in our task, we should not discard the skin and eyes under the transparent lenses of the spectacles. To solve the above problems, we manually picked out the occluded face images with incorrect segmentation

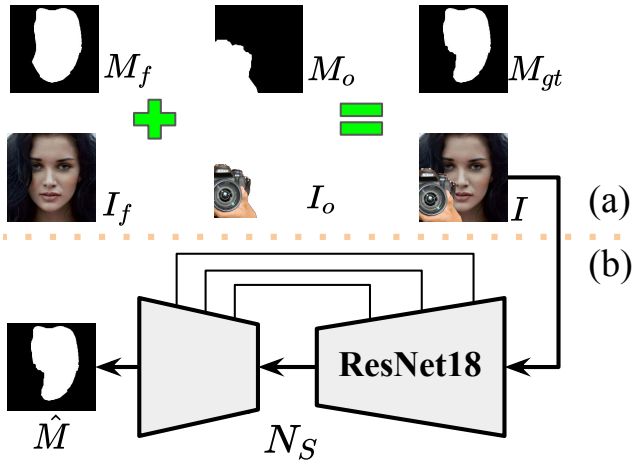

Figure 3. (a) Data augmentation method. I_f: face image, M_f: face mask of I_f, I_o: occlusion, M_o: occlusion mask, I: occluded image, M_{gt}: face mask of I. (b) Structure of the segmentation model N_S, $\hat{M}$: predicted face mask.

masks and labeled more than 1700 occlusions from them. We also collected about 300 occlusions from Google to get a more balanced occlusion type (e.g., masks, scarfs). In addition, we collected more than 800 texture patches covering every type we can think of, which will substitute for the original occlusion textures during training to produce more occlusion variations. Those occlusions are utilized for data augmentation, as illustrated in Figure 3 (a) and the following equation.

$$\begin{aligned} M_{gt} &= M_f \odot (1 - M_o), \\ I &= I_f \odot (1 - M_o) + I_o \odot M_o. \end{aligned} \tag{2}$$

Model Structure. N_S follows the classical U-Net structure [39], with ResNet18 [40] as encoder, while the decoder is a stack of convolutional blocks corresponding to each stage of the encoder. Each decoder block takes features from its previous block and its symmetric encoder stage through skip-connection. Since the target mask is binary, a single channel output for the decoder suffices.

Losses. The training is guided by the Dice loss [41] and the Binary Cross-Entropy loss (BCE):

$$L_{dice} = 1 - \frac{2 \sum \hat{M} \odot M_{gt}}{\sum \hat{M} + \sum M_{gt}}, \tag{3}$$

$$\begin{aligned} L_{bce} = -\frac{1}{WH} \sum \big(& M_{gt} \odot \log \hat{M} + \\ & (1 - M_{gt}) \odot \log(1 - \hat{M}) \big), \end{aligned} \tag{4}$$

where W and H represent the width and height of the mask, respectively. Further, we apply the "online hard example mining" (OHEM) [42] strategy to the BCE loss to make the training focus on hard examples, thus achieving more effective and efficient training.

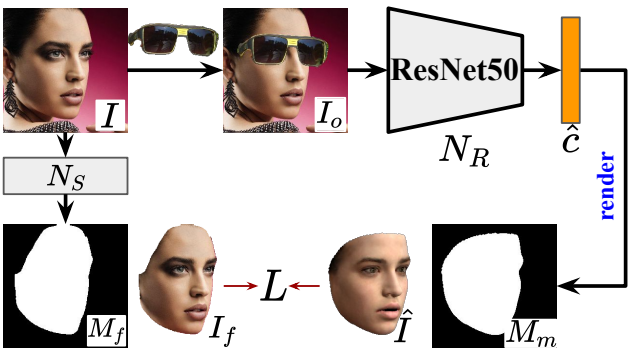

Figure 4. Overview of the 3D reconstruction module. I: original image, I_o: occluded image, $\hat{c}$: predicted 3D parameters, $\hat{I}$: reconstructed face, M_m: mask of $\hat{I}$, M_f: face mask of I, I_f: $I \odot M_f$

The proposed segmentation module achieves a precision of 0.981 and an IoU score of 0.954 on the validation dataset, which is acceptable considering the inaccurate edge region of the ground truth. By random checking the results, we surprisingly find that the predicted face region is often more accurate than the ground truth. We use this module to extract the face regions of the images in CelebA-HQ and FFHQ, which greatly facilitates the training of the other two modules, i.e., the reconstruction module N_R and the inpainting module N_G.

3.3. 3D Face Reconstruction

The reconstruction module is adapted from [37], the state-of-the-art 3D reconstruction method with a ResNet50 [40] as its backbone. Given a face image, it predicts a vector $c \in \mathbb{R}^{239}$, containing 6 translation and rotation parameters; 144 shape and 80 texture coefficients of the 3DMM; and nine illumination coefficients of the Spherical Harmonics [43, 44] model. With vector c, the face is reconstructed and further rendered to the image through a differentiable renderer.

Although [37] performs well on occlusion-free faces and even demonstrates certain robustness to small occlusions, it still cannot handle severe occlusions, e.g., sunglasses and masks. As shown in the second row of Figure 5, sunglasses make the reconstructed face have dark circles. In the third column, they even change the color of the reconstructed skin. Therefore, in our usage scenario, i.e., face de-occlusion, it is necessary to retrain an occlusion-robust 3D face reconstruction module. Now we present, as shown Figure 4, our training strategy.

Training data. The training is based on CelebA-HQ and FFHQ. Firstly, we use the pre-trained model of [37] to predict the reconstruction parameters of the training data, denoted as c_{gt}. Then, we filter out all the images with sun-

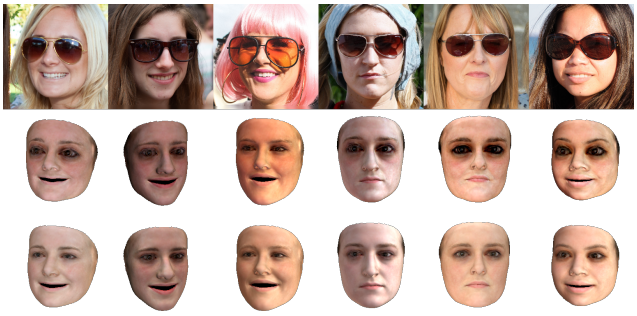

Figure 5. The first row shows the input images; the second row and the last row are reconstructed faces by [37] and ours, respectively.

glasses since their corresponding c_{gt} risk to be inaccurate. Next, we leverage the face segmentation module N_S to detect the face masks of the training data, denoted as M_f, for calculating the face-related losses. At last, as with the previous section, we randomly superimpose heavy occlusions to I during training to get occluded input images for N_R.

Losses. Since we already have c_{gt}, the most straightforward option is to use it as supervision; the loss is as:

$$L_{coef} = \frac{1}{N}\|\hat{c} - c_{gt}\|_1, \tag{5}$$

where $\hat{c}$ is the predicted coefficient, and N is its dimension.

Training solely with Equation 5 will result in inaccurate and non-discriminative results. Since it equally optimizes all the coefficients. However, different parts of $\hat{c}$ obviously have different impacts on the reconstruction result, e.g., the poses play a more critical role than the illumination coefficients. Therefore, we also leverage the pixel-wise and the perceptual [45] losses to reduce the discrepancy between the reconstructed face $\hat{I}$ and the real face I_f:

$$L_{pix} = \frac{1}{\sum M}\|\hat{I} \odot M - I_f \odot M\|_2, \tag{6}$$

$$L_{id} = 1 - \frac{\mathcal{F}(\hat{I})^T \cdot \mathcal{F}(I_f)}{\|\mathcal{F}(\hat{I})\| \cdot \|\mathcal{F}(I)\|}, \tag{7}$$

where M is the overlap of the original face mask M_f and the reconstruction mask M_m; $\mathcal{F}(\cdot)$ denotes the feature embedding function of a pre-trained face recognition model, and we use Arcface [46] here. We only extract the feature of I_f, rather than the entire image I, to avoid the noises introduced by the non-facial patterns.

To accelerate the training, we also use the landmark loss. Thanks to c_{gt}, we have ground truth 3D facial landmarks $\mathbf{q}$, thus eliminating the need for a 2D facial landmark detector as required in other methods. The loss is calculated as :

$$L_{ldmk} = \frac{1}{n_{pt}}\sum_{i=1}^{n_{pt}} \omega_n \|\hat{\mathbf{q}}_i - \mathbf{q}_i\|^2, \tag{8}$$

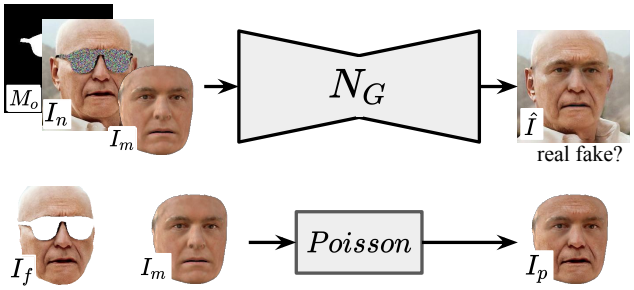

Figure 6. Overview of the inpainting module.

where n_{pt} denotes the number of landmarks, $\hat{\mathbf{q}}_i$ denotes the predicted 3D coordinates of the i-th facial landmark. The weights ω_n are set to 20 for the nose and inner mouth points and 1 for others.

As c_{gt} provides sufficient regularity, we discard the complex regularization terms in [37]. The overall loss is the weighted sum of the above losses:

$$L = L_{coef} + \lambda_{pix}L_{pix} + \lambda_{id}L_{id} + \lambda_{ldmk}L_{ldmk}, \tag{9}$$

where $\lambda_{pix} = 1.92$, $\lambda_{id} = 0.2$, $\lambda_{ldmk} = 1.6e^{-3}$.

Figure 5 compares our results with those of [37], demonstrating the effectiveness of the proposed training strategy in improving the occlusion-robustness of the model. A quantitative comparison is also performed: we use [37] and our model to reconstruct 1000 face images with synthetic occlusions, the L_{pix} of our model is 0.153, which is much less than 0.177 of [37].

3.4. Face Inpainting

With the N_S and N_R described above, we get the following information of an occluded face image: face mask M_f, face I_f, reconstructed face I_m, reconstructed face mask M_m. Based on M_f and M_m, Equation 1 further calculates the occlusion mask M_o. The goal of this section is to restore the missing textures indicated by M_o.

The generator N_G is constructed from stacked gated residual blocks [22] and follows the classical encoder-decoder structure; a self-attention [47] module is applied to the bottleneck features. As shown in Figure 6, the input is composed of three parts: the occlusion mask M_o, the face image with noised occlusion I_n, and the reconstructed face I_m. The output $\hat{I}$ is the de-occluded image. A VGG [48] shaped discriminator is further employed to increase the photorealism to $\hat{I}$.

We randomly superimpose occlusions to the training data and train N_G to recover the missing textures. One issue to note is that the input image itself always contains occlusions, such as bangs, and we do not have ground truth for these parts. So we use Poisson Blending [49] to seamlessly merge I_f and I_m and get I_p, which provides weak supervi-

sion for the inpainting task. The training is guided by the following losses:

Pixel-wise face loss.

$$L_{pix} = \frac{1}{\sum M} \|\hat{I} \odot M - I_f \odot M\|_1, \tag{10}$$

where $M = M_m \odot M_f$, limiting the loss to be calculated on the face region solely. This loss serves for recovering textures occluded by the synthetic occlusion and cannot recover the initially occluded textures.

SSIM loss. We use I_p to guide the generation of initially occluded textures. Although the Poisson blending result is visually pleasing, it changes the color of I_f, so we cannot simply apply L_1 or L_2 loss. Instead, we leverage the Structural Similarity loss (SSIM) [50], emphasizing the structural level discrepancy:

$$L_{sm} = \frac{-1}{\sum \bar{M}_m} SSIM(\hat{I} \odot \bar{M}_m, I_p \odot \bar{M}_m), \tag{11}$$

where $SSIM$ stands for the similarity mapping function (see Appendix for details); $\bar{M}_m$ is the eroded M_m, eliminating the edge effects in the similarity map. To further mitigate the impact of inaccurate color of I_p and make the loss focus on the missing textures, we apply the OHEM again as in Equation 4.

Background loss. Areas other than the face should remain unchanged:

$$L_{bg} = \frac{1}{\sum M_{bg}} \|\hat{I} \odot M_{bg} - I \odot M_{bg}\|, \tag{12}$$

where M_{bg} denotes the background mask, calculated by $1 - M_m$. We also erode M_bg to alleviate the edge effects.

Identity loss. The generated image should have the same identity as the original image, so we also leverage Equation 7 to add feature level identity constraints.

TV loss. To penalize the noises in $\hat{I}$, we adopt the total variation loss [51].

$$L_{tv} = \frac{1}{WHC} \|\nabla_x \hat{I}\|^2 + \|\nabla_y \hat{I}\|^2, \tag{13}$$

where W, H, C are the width, height, and the number of channels, respectively, $\nabla_{_}$ calculates the image gradient along a direction.

Adversarial loss. To further make $\hat{I}$ photorealistic, we employ the adversarial loss:

$$L_{adv} = -\mathbb{E}_{\hat{I}}[\log(D(\hat{I}))], \tag{14}$$

where $D(\cdot)$ denotes the output of the discriminator, which is the term to be maximized by the generator.

The global objective function of the generator can be summarized as follows:

$$\begin{aligned} L = &\lambda_{pix} L_{pix} + \lambda_{sm} L_{sm} + \lambda_{bg} L_{bg} + \\ &\lambda_{id} L_{id} + \lambda_{tv} L_{tv} + \lambda_{adv} L_{adv}, \end{aligned} \tag{15}$$

where the weights are empirically set to $\lambda_{pix} = 10$, $\lambda_{sm} = 5$, $\lambda_{bg} = 5$, $\lambda_{id} = 0.2$, $\lambda_{tv} = 0.1$, $\lambda_{adv} = 0.01$.

Discriminator loss. We use the BCE loss to train the discriminator, aiming to distinguish $\hat{I}$ from real images:

$$L_{adv} = \mathbb{E}_I[\log(D(I))] + \mathbb{E}_{\hat{I}}[\log(1 - D(\hat{I}))], \tag{16}$$

where I is the real image with less occlusion, selected from CelebA-HQ and FFHQ according to the overlap rate of its corresponding M_m and M_f.

4. Experiments

The proposed method can effectively remove the occlusions from face images. To demonstrate the power of our model, we qualitatively compare our results with state-of-the-art methods, including those for face de-occlusion and for image inpainting. Quantitative experiments are conducted in two aspects as well, namely reconstruction ability and identity recovery ability.

4.1. Implementation Details

We follow an incremental training strategy: 1) We train the segmentation module N_S. 2) we train the reconstruction module N_R based on N_S. 3) We train the inpainting module N_G based on the result of the first two. The training is performed on two public datasets, CelebAMask-HQ [38] and FFHQ [52], and our manually labeled/collected occlusions and textures. All the images are aligned with facial landmarks detected by [53] and cropped to 256×256.

The training of N_S consists of two rounds: In the first round, we train on 300,000 images of CelebAMask-HQ with our manually labeled occlusions to obtain a coarse model. In the second round, we apply the coarse model to both CelebAMask-HQ and FFHQ [54], select 500 hard examples from the results, relabel and add them to the occlusion dataset to retrain a more accurate model. Each round is trained for 30 epochs with a batch size of 16 and a learning rate of $1e^{-4}$. The training takes about two hours on two Nvidia GTX 1080 GPUs. The 3D reconstruction module and the face inpainting modules are trained on 100,000 images of CelebAMask-HQ and FFHQ, with an initial learning rate of $1e^{-4}$. For the reconstruction module N_R, we train 50 epochs with a batch size of 16, and for the inpainting module N_G, we train 80 epochs with a batch size of 4. The learning rates of both modules are dropped halfway through the training with ratios of 0.2 and 0.1, respectively. All three modules are optimized using Adam [55] with a weight decay of 0. Betas are set to $[0.9, 0.999]$ for N_S and N_R, $[0.5, 0.999]$ for N_G. It takes about 80 hours to train N_R and 40 hours to train N_G on two Titan X Pascal GPUs.

4.2. Qualitative Results

The method closest to ours is FD3D [26]; we both leverage 3D face reconstruction for face de-occlusion. Unfortu-

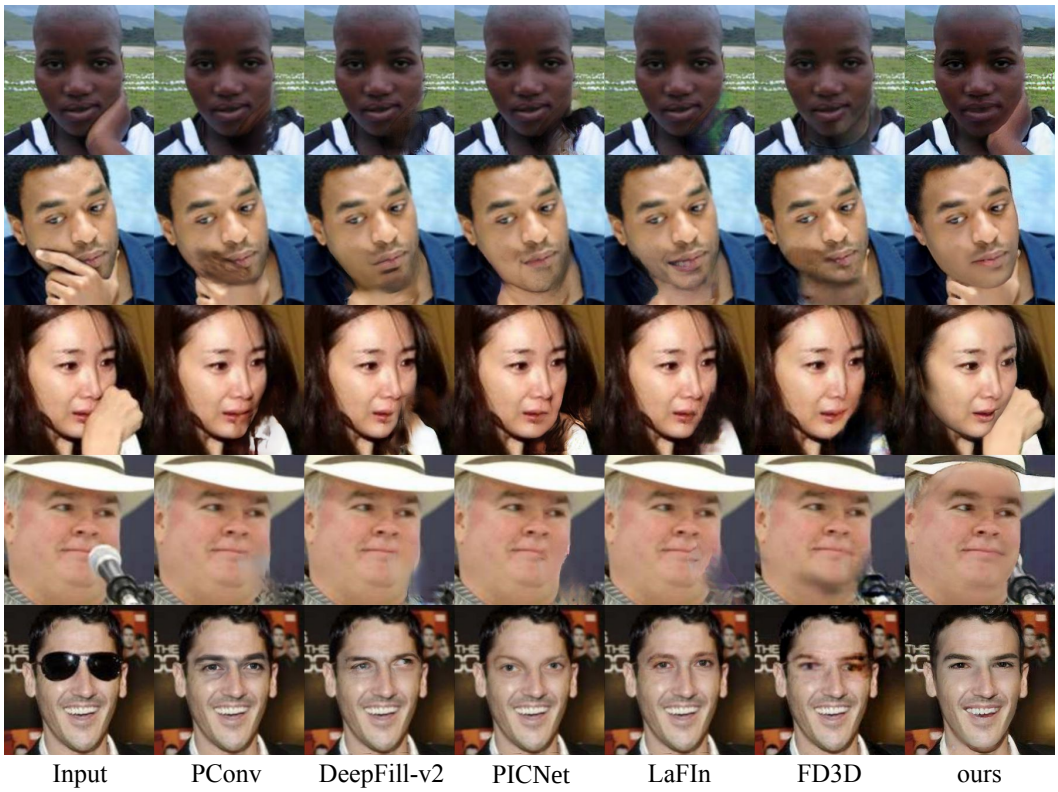

Figure 7. De-occlusion results compared with state-of-the-art methods. Input images are aligned and cropped according to the rules of each method, and we re-paste the outputs back to facilitate comparison.

nately, they do not release their code, so we use images from their paper to conduct the qualitative evaluation. Figure 7 compares our method with FD3D and several other publicly available image inpainting methods, including DeepFill-v2 [22], PConv [21], PICNet [56], and LaFIn [34]. The FD3D results are taken from their paper, while the rest are generated from their official implementations with models pre-trained on face images. We provide those inpainting methods with manually labeled occlusion masks for a fair comparison. One issue to note is that when FD3D compares with other methods in their paper, the images are not pre-aligned according to the rules of the corresponding method, resulting in questionable inferior results.

As shown in Figure 7, our method generally outperforms FD3D and other image inpainting methods, especially on face images with large yaw angles. We argue that the image inpainting methods are purely based on statistical learning, thus highly dependent on the training data distribution. When a few face images with large yaw angles occur, these methods record dramatic performance degradation. Although FD3D does not require a manual mask for the de-occlusion, they apparently can only handle limited types of occlusions: the bangs in the third row and the hat in the fourth row are not identified as occlusions. In addition, FD3D produces low-quality face images with blurred boundaries and unreal textures, as evidenced in the images from the second to the last row. We believe this is due to the following reasons: 1) they do not explicitly detect the occlusions and mask them out, causing the generator to only cope with the occlusions it has ever seen while failing to de-occlude arbitrary occlusions; 2) they use a very coarse 3D reconstruction method, which cannot provide correct and effective face prior to the generator.

4.3. Quantitative Results

This section evaluates the proposed method in two aspects: the reconstruction ability and the identity recovery ability. We synthesize 1000 sunglasses-occluded face images and use DeepFill-v2, PICNet, LaFIn, and our method to recover the initial images. Since the image inpainting methods only synthesize the missing region, the reconstruction ability is evaluated only in that region, with the following metrics: L1 loss, SSIM [50] score, and PSNR score. The identity recovery ability is evaluated by the cosine similarity of the features extracted by the pre-trained ArcFace [46]. Results are reported in Table 1.

As can be seen, our method outperforms the state-of-the-art image inpainting methods across all listed metrics. It

Method	$L_1\downarrow$	SSIM $\uparrow$	PSNR$\uparrow$	ID$\uparrow$
LaFIn	0.067	0.607	27.944	0.639
PICNet	0.0654	0.604	28.137	0.615
DeepFill	0.065	0.604	28.105	0.659
w/o SSIM	0.065	0.619	28.491	0.665
Ours	**0.062**	**0.623**	**28.902**	**0.690**

Table 1. Comparison of the proposed method with state-of-the-art image inpainting methods.

performs particularly well in recovering the identity, much exceeding its comparators. We attribute this mainly to the occlusion-robust face reconstruction module.

4.4. Ablation Study

The three modules in our framework are interdependent, making it impossible to remove one for ablation study. Instead, we mainly focus on the SSIM loss of the inpainting module, as the reconstruction and identity losses are conventional practices.

As Table 1 shows, without the SSIM loss, the model degrades in all aspects of metrics (nevertheless, benefiting from the 3D prior, it still outperforms others). Moreover, the initially occluded textures are now only supervised by the adversarial loss; thus, the results are not guaranteed to be occlusion-free (see Appendix for examples).

4.5. Discussion

Limitation. The proposed method focuses on removing the occlusions within the 3D reconstructed face mask. For occlusions spanning the face and the background, it can only rigidly remove the parts above the face without producing a smooth transition, creating an unrealistic scene where the face seems to hover above the occlusions, as Figure 8 shows. We have tried to create a gap between the face area and the background through mask erosion, and the gap is solely supervised by the adversarial loss during training. This trick alleviates the boundary effect to some extent but still cannot cope with large-sized occlusions.

Social impacts. Our method helps pre-process face images to avoid the negative impact of occlusions such as hair, glasses, hands, etc., on downstream tasks (e.g., fine-grained 3D face reconstruction, face recognition). Despite the benefits it brings, it also risks violating human privacy. To quantify this risk, we further conduct experiments to analyze the identity recovery ability across different types of occlusions. Figure 9 shows that the proposed method performs well on sunglasses-occluded faces; the performance decreases for the mask-type occlusion; and when combining the mask with sunglasses, the method can no longer recover the face identity. The above observation proves that the method is controllable for privacy violations. People's identities can be safely protected when they simultaneously wear sunglasses and masks.

Figure 8. Limitation of the proposed method.

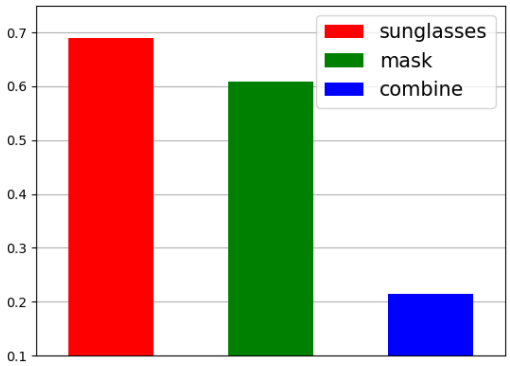

Figure 9. Identity recovery ability for different types of occlusion.

5. Conclusion

This work proposes a segmentation-3D reconstruction-guided facial image de-occlusion method that automatically removes all types of occlusions. We first analyze the limitations of current face de-occlusion methods: they either require a manually labeled mask or can only handle a limited number of occlusion types, which mainly stems from the vast diversity of possible occlusions. Our key innovation is bypassing the segmentation of infinite occlusions and instead segmenting the face regions, which is much easier. We manually labeled a large face occlusion dataset, based on which we trained a face segmentation module N_S and an occlusion-robust 3D reconstruction module N_G. Given an occluded face image, N_S and N_G work collaboratively to mask the occlusions and provide beneficial priors to guide the subsequent face inpainting module. Qualitative and quantitative evaluations demonstrate the superiority of the proposed method.

References

[1] J.-J. Lv, X.-H. Shao, J.-S. Huang, X.-D. Zhou, and X. Zhou, "Data augmentation for face recognition," *Neurocomputing*, vol. 230, pp. 184–196, 2017. 1

[2] D. S. Trigueros, L. Meng, and M. Hartnett, "Enhancing convolutional neural networks for face recognition with occlu-

sion maps and batch triplet loss," *Image and Vision Computing*, vol. 79, pp. 99–108, 2018. 1

[3] Y. Xia, B. Zhang, and F. Coenen, "Face occlusion detection based on multi-task convolution neural network," in *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp. 375–379, IEEE, 2015. 1

[4] W. Wan and J. Chen, "Occlusion robust face recognition based on mask learning," in *2017 IEEE international conference on image processing (ICIP)*, pp. 3795–3799, IEEE, 2017. 1

[5] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential siamese network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 773–782, 2019. 1

[6] H. Ben Fredj, S. Bouguezzi, and C. Souani, "Face recognition in unconstrained environment with cnn," *The Visual Computer*, vol. 37, no. 2, pp. 217–226, 2021. 1

[7] L. He, H. Li, Q. Zhang, Z. Sun, and Z. He, "Multiscale representation for partial face recognition under near infrared illumination," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–7, IEEE, 2016. 1

[8] Y. Wu and Q. Ji, "Robust facial landmark detection under significant head poses and occlusion," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3658–3666, 2015. 1

[9] L. He, H. Li, Q. Zhang, and Z. Sun, "Dynamic feature learning for partial face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7054–7063, 2018. 1

[10] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009. 1

[11] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991. 1

[12] C. Wu, C. Liu, H.-Y. Shum, Y.-Q. Xy, and Z. Zhang, "Automatic eyeglasses removal from face images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 3, pp. 322–336, 2004. 1

[13] J.-S. Park, Y. H. Oh, S. C. Ahn, and S.-W. Lee, "Glasses removal from facial image using recursive error compensation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 5, pp. 805–811, 2005. 1

[14] Z.-M. Wang and J.-H. Tao, "Reconstruction of partially occluded face by fast recursive pca," in *2007 International Conference on Computational Intelligence and Security Workshops (CISW 2007)*, pp. 304–307, IEEE, 2007. 1

[15] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006. 1

[16] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Advances in neural information processing systems*, pp. 341–349, 2012. 1

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014. 1

[18] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016. 1, 2

[19] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3911–3919, 2017. 2

[20] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5505–5514, 2018. 2

[21] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 85–100, 2018. 2, 7

[22] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4471–4480, 2019. 2, 5, 7

[23] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, "Robust lstm-autoencoders for face de-occlusion in the wild," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 778–790, 2017. 2

[24] J. Cai, H. Han, J. Cui, J. Chen, L. Liu, and S. K. Zhou, "Semi-supervised natural face de-occlusion," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1044–1057, 2020. 2

[25] J. Dong, L. Zhang, H. Zhang, and W. Liu, "Occlusion-aware gan for face de-occlusion in the wild," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2020. 2

[26] X. Yuan and I. K. Park, "Face de-occlusion using 3d morphable model and generative adversarial network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10062–10071, 2019. 2, 6

[27] Y.-H. Lee and S.-H. Lai, "Byeglassesgan: Identity preserving eyeglasses removal for face images," in *European Conference on Computer Vision*, pp. 243–258, Springer, 2020. 2

[28] B. Hu, Z. Zheng, P. Liu, W. Yang, and M. Ren, "Unsupervised eyeglasses removal in the wild," *IEEE Transactions on Cybernetics*, 2020. 2

[29] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*, pp. 7354–7363, PMLR, 2019. 2

[30] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019. 2

[31] Y. A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim, "Unsupervised attention-guided image to image translation," *arXiv preprint arXiv:1806.02311*, 2018. 2

[32] Y. Lei, W. Du, and Q. Hu, "Face sketch-to-photo transformation with multi-scale self-attention gan," *Neurocomputing*, vol. 396, pp. 13–23, 2020. 2

[33] C. Zheng, T.-J. Cham, and J. Cai, "Pluralistic image completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1438–1447, 2019. 2

[34] Y. Yang and X. Guo, "Generative landmark guided face inpainting," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 14–26, Springer, 2020. 2, 7

[35] N. U. Din, K. Javed, S. Bae, and J. Yi, "A novel gan-based network for unmasking of masked face," *IEEE Access*, vol. 8, pp. 44276–44287, 2020. 2

[36] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 187–194, 1999. 2

[37] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. 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*, pp. 0–0, 2019. 3, 4, 5

[38] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 6

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015. 4

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 4

[41] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE, 2016. 4

[42] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769, 2016. 4

[43] R. Ramamoorthi and P. Hanrahan, "An efficient representation for irradiance environment maps," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 497–500, 2001. 4

[44] R. Ramamoorthi and P. Hanrahan, "A signal-processing framework for inverse rendering," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 117–128, 2001. 4

[45] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, pp. 694–711, Springer, 2016. 5

[46] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019. 5, 7

[47] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018. 5

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 5

[49] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *ACM SIGGRAPH 2003 Papers*, pp. 313–318, IEEE, 2003. 5

[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. 6, 7

[51] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196, 2015. 6

[52] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017. 6

[53] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017. 6

[54] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019. 6

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 6

[56] C. Zheng, T.-J. Cham, and J. Cai, "Pluralistic free-from image completion," *International Journal of Computer Vision*, pp. 1–20, 2021. 7