Facial Structure Guided GAN for Identity-preserved Face Image De-occlusion

Yiu-ming Cheung
ymc@comp.hkbu.edu.hk
Department of Computer Science,
Hong Kong Baptist University
Hong Kong, China

Mengke Li
csmkli@comp.hkbu.edu.hk
Department of Computer Science,
Hong Kong Baptist University
Hong Kong, China

Rong Zou
rongzou@comp.hkbu.edu.hk
Department of Computer Science,
Hong Kong Baptist University
Hong Kong, China

ABSTRACT

In some practical scenarios, such as video surveillance and personal identification, we often have to address the recognition problem of occluded faces, where content replacement by serious occlusion with non-face objects always produces partial appearance and ambiguous representation. Under the circumstances, the performance of face recognition algorithms will often deteriorate to a certain degree. In this paper, we therefore address this problem by removing occlusions on face images and present a new two-stage Facial Structure Guided Generative Adversarial Network (FSG-GAN). In Stage I of the FSG-GAN, the variational auto-encoder is used to predict the facial structure. In Stage II, the predicted facial structure and the occluded image are concatenated and fed into a generative adversarial network (GAN) based model to synthesize the de-occlusion face image. In this way, the facial structure knowledge can be transferred to the synthesis network. Especially, in order to enable the occluded face image to be perceived well, the generator in the GAN based synthesis network utilizes the hybrid dilated convolution modules to extend the receptive field. Furthermore, aiming at further eliminating the appearance ambiguity as well as unnatural texture, a multi-receptive fields discriminator is proposed to utilize the features from different levels. Experiments on the benchmark datasets show the efficacy of the proposed FSG-GAN.

CCS CONCEPTS

• **Computing methodologies** → **Object recognition**; **Object identification**; **Reconstruction**.

KEYWORDS

face de-occlusion; partial face recognition; generative adversarial networks

ACM Reference Format:
Yiu-ming Cheung, Mengke Li, and Rong Zou. 2021. Facial Structure Guided GAN for Identity-preserved Face Image De-occlusion. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21), August 21–24, 2021, Taipei, Taiwan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3460426.3463642

ICMR '21, August 21–24, 2021, Taipei, Taiwan

ACM ISBN 978-1-4503-8463-6/21/08…$15.00
https://doi.org/10.1145/3460426.3463642

1 INTRODUCTION

Face images in real-world scenarios like surveillance and forensics often suffer from different kinds of occlusions, as shown in Figure 1. Such occlusions often cause serious information scarcity, especially for the large occluded area. Subsequently, it severely reduces the reliability of the most advanced recognition algorithms, which will lead to crucial surveillance failure if one deliberately hides the face. This makes some tasks, e.g. face identification and recognition [34], face parsing [26], lip tracking and contour extraction [9, 28] becomes more challenging. Therefore, face image de-occlusion is a problem of both academic and practical importance.

Figure 1: Examples of occluded face images, where subfigure (a)-(c) are the face images occluded by mask[1], sunglasses[2] and car sun visor[3], respectively.

Recently, a number of generative models [41, 51] have been proposed to generate visually realistic images, but they ignore preserving the identity information. Furthermore, another related algorithm towards the face image de-occlusion, namely image completion [2], only works well under constrained occlusion shapes [5, 25, 48], or need masks of occlusion objects [27, 49]. Recently, Dong et al. [12] have proposed Occlusion-Aware GAN (OA-GAN) to address the arbitrary facial occlusions removing problem. This method successfully removes small occlusion, but meanwhile generating inconsistent texture when the occluded area is large. As far as we know, Image de-occlusion problem is still challenging due to the following four reasons:

1) No preservation of the identity information: Existing generative methods like Cycle-GAN [51] and DeepFill [49] can only transform the holistic style without preservation of the identity information, which is unfavorable for face recognition.

[1]The image comes from: https://www.cna.com.tw/news/firstnews/202007210272.aspx
[2]The image comes from: https://kknews.cc/world/xjj5llg.html
[3]The image comes from: http://gd.sina.com.cn/yj/social/2015-11-24/102825555.html

2) Occlusion mask unavailable: Methods like DeepFill and OA-GAN can fill in the missing area with semantic reasonable content. However, DeepFill requires a given binary mask of the occlusion part to tell the model where to fill in. Also, OA-GAN needs the image of the occlusion object as ground truth to obtain the "occlusion awareness".
3) Large occluded area: Even though de-occlusion methods like OA-GAN perform well for the small occlusion, the recovered result through OA-GAN lacks overall consistency when the occluded area is large.
4) New identities: The aforementioned methods divide the training and testing datasets based on images rather than identities. That is, the models may have seen a face image of the target person. For the occluded face images of a new identity whose photos are not included in the training dataset, the de-occlusion results from the existing methods will become deteriorate.

To this end, inspired by the amodal perception mechanism in human visual system [31], we propose a two-stage Facial Structure Guided GAN (FSG-GAN) to perform large area occluded face image de-occlusion without the occlusion mask, meanwhile preserving the identity information. Furthermore, the face images of the target person for testing are not essentially included in the training set. The model consists of two stages, i.e. facial structure prediction and de-occluded image synthesis. As the human face's topological structure of different individuals is similar, in Stage I, we therefore use the variational auto-encoder (VAE) [22] to predict the overall facial structure that can provide background knowledge for the subsequent de-occluded image synthesis. In Stage II, the predicted structure and the occluded image are fed into the GAN-based synthesis network to obtain the de-occluded image. In order to enable the occluded face to be perceived well, we extend the receptive fields (RFs) of the synthesis network by introducing Hybrid Dilated Convolution (HDC) [45] modules. Ledig et al. [24] have exploited perceptual loss and gained an effective performance boost recently. To maintain the identity information and benefit the subsequent recognition, we improve this perceptual loss by using the cosine distance to replace the originally used Euclidean distance. A multi-receptive fields discriminator is proposed to further eliminate the appearance ambiguity and unnatural texture. To evaluate the proposed model, we synthesize an occluded face image dataset based on CelebFaces Attributes (CelebA) [29] and Labeled Faces in the Wild (LFW) [18]. The experimental results validate the compelling effectiveness of the proposed method.

The main contributions of this paper are highlighted below:

1) We propose a novel framework to remove different kinds of occlusions on face images.
2) We introduce the HDC modules to the proposed network to improve the holistic consistency of de-occluded face images.
3) We propose a multi-receptive fields discriminator that discriminates the different RFs features to further enhance the detail of de-occluded face images.
4) We build a synthetic occluded face image dataset[1], on which occlusions are semantically placed with reference to face landmarks.

2 RELATED WORK

2.1 Image Completion

Image completion (also called *inpainting*) aims to fill in a missing region of an image automatically with visually plausible pixels. Traditional methods like diffusion-based methods [6, 15, 39] and patch-based methods [1, 10]. The former distributes the external information along the contour normal into the missing portion, while the latter copying patches information from a similar area of an image (or a set of images) can handle simple stationary texture image well. Recently, benefiting from generative models, a number of image completion methods [5, 25, 48], which can deal with complex images, have been proposed. These methods exploit the encoder-decoder based network combined with the reconstruction loss and the adversarial loss to recover the missing contents. However, these methods only work well under specifically restricted occlusions like rectangular shape, random noise, etc., leaving more complicated practical occlusions unresolved. Recently, Yu et al. [49] have proposed DeepFill which fills in multiple irregular shaped holes at arbitrary locations in an image. This model uses a coarse network to generate missing contents, and meanwhile using another refinement network with a contextual attention module to enhance image details and spatial coherency. It generally requires the binary occlusion mask during testing. Furthermore, image completion methods mainly focus on making the generated faces vividly, but neglect to maintain the personal identity information.

2.2 Frontalization and Face Image De-occlusion

Frontalization focuses on synthesizing the frontal view of a given side face image. Frontalization and face image de-occlusion both aim to increase the performance of face recognition algorithms. Therefore, retaining the subject identity is essential. Traditional frontalization methods use 2D/3D surface texture warping [3, 14, 17], or landmark localization methods [38]. Conventional face image de-occlusion algorithms address the problem with sparse coding techniques [33, 46] or filtering methods [30, 35]. Thanks to the revolution of deep neural networks, lots of face frontalization and de-occlusion works [12, 19, 50] have been researched and achieved outstanding performances. For example, Huang et al. [19] have proposed a frontal view synthesis method called TP-GAN, which consists of a global and a local generator. By several regularization functions like the symmetry loss and the total variation loss, the model can generate photorealistic frontal view image. Moreover, Dong et al. [12] have proposed a two-stage Occlusion-Aware GAN (OA-GAN) that firstly segments the occluded part of the face image with a GAN based model and then removes this occlusion with another GAN model. This method can remove real-world occlusions and generate visually realistic images, but it needs to use the occlusion object as the ground truth to indicate the occluded part during the first stage of training. Nevertheless, TP-GAN and OA-GAN still need the basic facial features. These methods are not competent when some facial organs, e.g. eyes or mouth, are totally unseen.

[1]The full dataset can be found in https://drive.google.com/drive/folders/1ISmIMmmpEVFTi8Xl2aiGR8DUBjjOEHMl?usp=sharing

3 THE PROPOSED METHOD

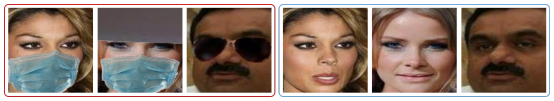

Figure 2: Examples of the training set $\{X, Y\}$, where (a) shows the occluded face images $\{x_i\} \in X$, and (b) is the original occlusion-free face images $\{y_i\} \in Y$.

In this section, we will propose the FSG-GAN for face image de-occlusion. Specifically, the proposed method is intended to recover the occluded area and then the output image is utilized to recognize the right person. Human beings can be conscious of the part of a person that is occluded behind an object. How we represent these occluded parts of perceived objects is the problem of amodal perception. [32] pointed out that amodal perception heavily relies on our background knowledge of the appearance of the (possibly) occluded part of the object. Using the most quoted example in amodal perception, namely the cat behind the picket fence, it is natural that a person who has never seen a cat will have difficulty in representing its occluded tail behind the fence. In our work, we utilize VAE to construct the background knowledge, then use a GAN-based synthesis network to obtain the final de-occlusion result. As shown in Figure 2, let $\{X, Y\}$ denote the training set, where X represents the occluded face image set and Y represents the corresponding original occlusion-free face image set. For each $x \in X$ of the target person, we can obtain the predicted structure g that estimates $y \in Y$ as close as possible. Then, g is used as the background knowledge to de-occlude the face image of the target person, and eventually get the output $\hat{y}$.

3.1 The Framework of FSG-GAN

The overall framework of our proposed FSG-GAN is shown in Figure 3, which is composed of two stages. We firstly predict the holistic facial structure of the occluded image by VAE in Stage I. VAE makes a strong assumption that the posterior is approximately factorial and predictable with a neural network [4]. These assumptions constrain the expressive ability of the model, thus resulting in blurry results. Even though a lot of previous works [16, 36, 41] increased the expressiveness of the approximate prior and posterior, and promising results were obtained, they take a long time to train. Our model aims to take advantage of the fast and tractable training of VAE. Therefore, we use the generated blurry result by VAE (denoted as V) as the holistic facial structure to provide background knowledge for the following de-occlusion procedure. The predicted structure g can be denoted as:

$$g = V(x). \tag{1}$$

Then, a U-Net [37] structure generator (G) is applied to get the de-occluded face image in Stage II. The result obtained by V provides the geometric background knowledge of a human face. G can efficiently remove the occlusion even when the occluded area is large with the guidance of the obtained geometric background knowledge. The de-occluded face image $\hat{y}$ is denoted as:

$$\hat{y} = G(x, g). \tag{2}$$

To further enhance the de-occlusion results, an adversarial structure is utilized. Different from the previous work [49] that exploits a two pathway discriminator to fuse global and local critics, our discriminator D gives the final output that is a soft version of the weighted estimation of different RFs features.

3.2 Facial Structure Prediction

VAE assumes that both the prior and posterior are Gaussian distributions. The performance of such model is constrained by the restrictive mean field approximation to the intractable posterior distribution [40], and thus resulting in blurred generated samples. In this paper, we take the advantage of fast learning complicated distributions by VAE to obtain the complete face image distributions. We assume that the occluded image domain has intersection with the complete image in the latent space. The purpose of the encoder of V is to map the occluded image domain into this intersection in the latent space, and then the decoder is to find the corresponding complete face image. Some features are occluded, which increases the uncertainty of the corresponding complete face image. Therefore, V outputs a blurry image. We use the blurry image as the overall structure to provide background knowledge for Stage II. To train V, we optimize the variational lower bound L_b on $\log p(y)$:

$$\begin{aligned}
\log p(g) &= \int_z q(z|x) \log p(g) \mathrm{d}z \\
&= \int_z q(z|x) \log \left(\frac{p(z,g)}{q(z|x)} \frac{q(z|x)}{p(z|g)} \right) \mathrm{d}z \\
&= \int_z q(z|x) \log \left(\frac{p(z,g)}{q(z|x)} \right) \mathrm{d}z + \mathrm{KL}\left(q(z|x) \| p(z|g)\right) \\
&= L_b + \mathrm{KL}\left(q(z|x) \| p(z|g)\right).
\end{aligned} \tag{3}$$

The KL(.) term is always greater or equal to zero. Therefore, the objective of V is to maximize the variational lower bound L_b. The loss function L_{VAE} for V is:

$$\begin{aligned}
L_{VAE} &= -\int_z q(z|x) \log \left(p(g|z)\right) \mathrm{d}z - \int_z q(z|x) \log \left(\frac{p(z)}{q(z|x)} \right) \mathrm{d}z \\
&= \mathbb{E}_{z \sim q(z|x)}[-log p(g|z)] + \mathrm{KL}\left(q(z|x) \| p(z)\right),
\end{aligned} \tag{4}$$

where $q(z|x) = z \sim \mathcal{N}(\mu_e(x), \sigma_e(x))$, $p(g|z) = g \sim \mathcal{N}(\mu_d(z), \sigma_d(z))$ and the prior $p(z)$ is set as an isotropic unit Gaussian $z \sim \mathcal{N}(0, 1)$. (μ_e, σ_e) and (μ_d, σ_d) are obtained by the encoder and decoder blocks in V, respectively. After optimization, the overall structure g of the occluded image x can be obtained by the trained V.

3.3 De-occluded Face Image Synthesis

The main framework of our de-occluded image synthesis model consists of a U-Net structure generator G and a multi-RFs discriminator D. The generator takes the occluded image x and the overall structure g as input. The discriminator D shares the same structure with the encoder blocks in VAE, and is then followed by three shallow nets that attempt to determine the real or fake image from the different levels.

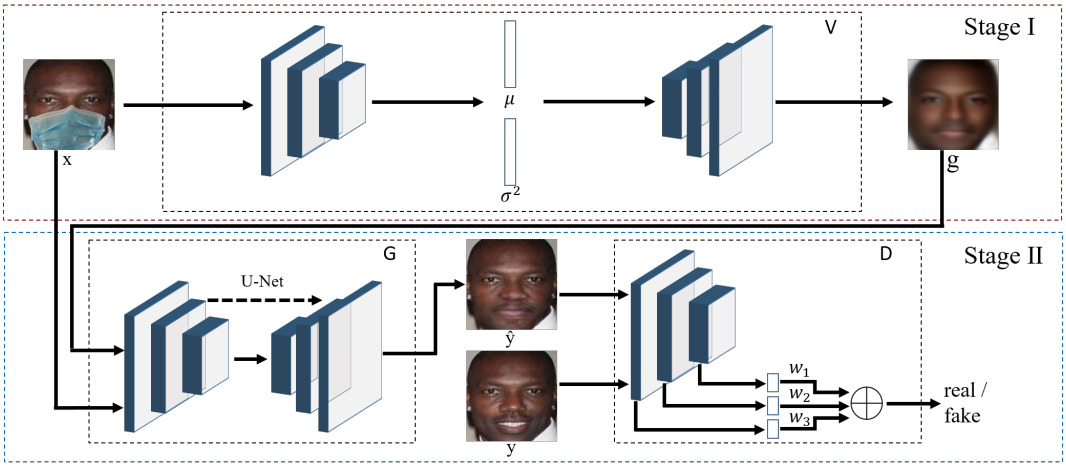

Figure 3: An overview of the proposed FSG-GAN. The training process consists of two stages. Stage I utilizes the VAE (V) to predict the overall facial structure, and Stage II takes the U-Net structure generator (G) and the proposed multi-RFs discriminator (D). The final output of D is a softmax weighted sum of the discriminant results from the different layers.

3.3.1 De-occluded Image Generator. G takes a U-Net structure where skip connections are utilized to fuse the multi-level features. The encoder-decoder structure has been successfully used, e.g. see [19, 20]. However, the generated images suffer from problems like inconsistency in overall skin tone and unnatural recovered regions. We therefore introduce the HDC module to encoder blocks to solve such problems. Dilated (atrous) convolution has been widely used in image semantic segmentation [7, 8], speech synthesis [42], machine translation [21], etc. It can enlarge the RFs of the model without reducing the size of the feature map to aggregate global information. Wang et al. [45] have proposed HDC to further improve the model by utilizing a range of dilation rates and concatenating them serially, which can effectively alleviate the gridding issue caused by the standard dilated convolution. We take advantage of HDC to make the output image globally consistent without losing the details.

Structure. Figure 4 shows the details of the generator, which exploits the U-Net structure consisting of three encoder blocks and three decoder blocks. Each encoder block contains a down-sampling module, an HDC module and a residual module. The HDC module parallels one 1×1 convolution and three dilated convolutions with the different dilated rates capturing the multi-scale information. The kernel size of the dilated convolutions is 3×3 and the dilated rates are set at $[1, 2, 5]$, respectively. The batch normalization (BN) and swish activation suggested by NVAE [41] are used in each module. In this way, the RFs of the three encoder blocks are a quarter, a half and the entire image, respectively. The decoder block has an up-sampling module that contains an up-sampling unit and a convolution layer. The skip connection is used from the corresponding symmetrical block.

Loss function. L1 norm is adopted as the metric of the distance between the de-occluded image and the ground truth image. As a result, the pixel-wise loss L_{pw} takes the form:

$$L_{pw}(G) = \mathbb{E}_{(x,y)\sim p_d(x,y)}\left[\|y - G(x,g)\|_1\right]. \quad (5)$$

This L1 formed pixel-wise loss forces the generator to output a sharper image than the L2 norm [20].

Preserving crucial identity information while synthesizing the de-occluded face image is essential for the subsequent application like face recognition. The perceptual loss originally proposed by Ledig et al. [24] can maintain perceptually relevant similarity in super-resolution. We exploit this perceptual loss to help the model preserve the identity similarity. Differently, we replace the originally used VGG19 with LightCNN [47] because LightCNN is specifically for face images and is much smaller. In addition, the previous works [43, 44] have shown that identity information is only related to the angles of the deep features. Therefore, we define the identity-preserve loss L_{ip} on the LightCNN features by the cosine distance:

$$L_{ip}(G) = \mathbb{E}_{(x,y)\sim p_d(x,y)}\left[1 - \frac{F(y)\cdot F(G(x,g))}{\|F(y)\|_2 \, \|F(G(x,g))\|_2}\right], \quad (6)$$

where F represents the pre-trained LightCNN model.

The adversarial loss L_{adv} is used to further enhance the realism of the de-occluded image:

$$L_{adv}(G) = \mathbb{E}_{x\sim p_d(x)}\Big[-\log\Big(D\big(G(x,g)\big)\Big)\Big], \quad (7)$$

where D is the proposed multi-RFs discriminator that will be discussed in detail in the following subsection 3.3.2.

The overall loss function for G is a weighted sum of the previously defined losses:

$$L_{total}(G) = \lambda_{pw}L_{pw}(G) + \lambda_{ip}L_{ip}(G) + \lambda_{adv}L_{adv}(G). \quad (8)$$

3.3.2 Multi-RFs Discriminator. The L1 and L2 norm losses favor blurry results on image generation because these losses fail to capture high-frequencies in many cases. In the literature, Isola et al. [20] have proposed Markovian discriminator to encourage generator model high-frequencies by determining $N \times N$ real or fake patches in an image. In our face image de-occlusion problem, the global consistency and local details are both important. Different from the previous work like DeepFill [49] that introduces global

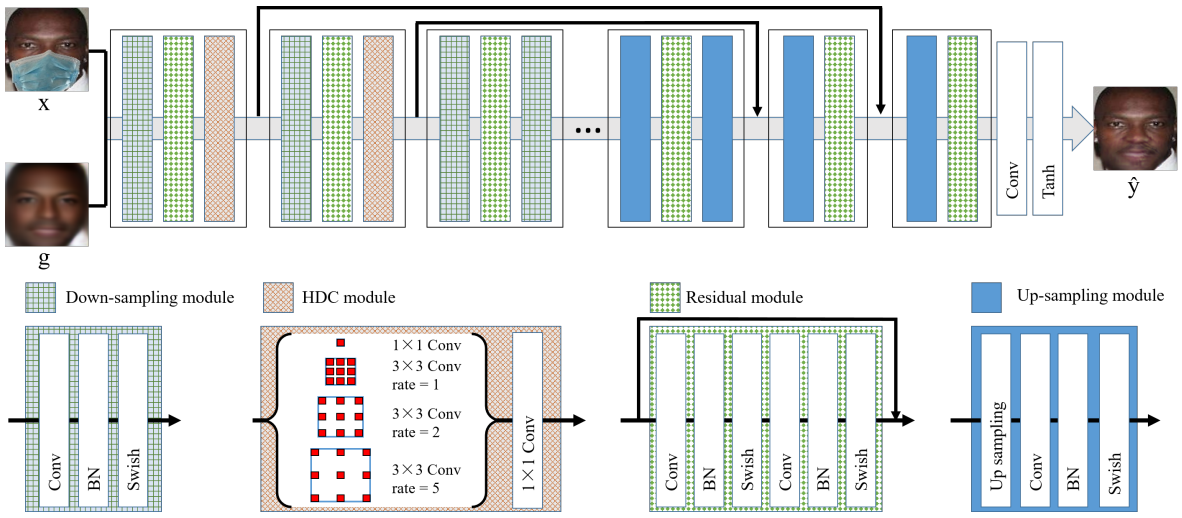

Figure 4: The detailed structure of the generator and each module. The generator takes the predicted structure image g by VAE and its corresponding occluded image x as input.

and local discriminators, we improve the Markovian discriminator by determining the features from the different layers. It is well known that the RF in shallow network is small and will increase as the network deepens. Discriminating different level features that have different RFs can achieve the goal of fusing global consistency and local details. For the i-th layer features, we have $N_i \times N_i$-patch responses. The i-th layer output d_i is an average of all these responses. The ultimate output of D is a weighted sum of all layers' output, where the weights are calculated by a softmax function.

Structure. The feature extraction module in the discriminator shares the weight with the encoder in VAE to reduce the number of training parameters. The feature maps from different layers are fed into discriminator modules which all consist of a convolution, a nonlinear activation, and a linear unit followed by a sigmoid.

Loss function. We have I outputs of different RFs feature maps. The ultimate output d_u of D is calculated by:

$$d_u = \sum_{i}^{I} w_i d_i, \tag{9}$$

where d_i represents the response of the i-th layer. We use a soft version parameterized by λ of the three classical Pythagorean means [13] to calculate the weight w_i:

$$w_i = \frac{e^{\lambda d_i}}{\sum_j^I e^{\lambda d_j}}, \tag{10}$$

where $\lambda \geq 0$. d_u is the arithmetic mean of I outputs when $\lambda = 0$ and the max value when $\lambda \to \infty$. We take $\lambda = 1$, then Eq. (9) becomes a softmax that is differentiable. The Minimax objective function is utilized for D:

$$\begin{aligned} L_{adv}(D) =& \mathbb{E}_{y\sim p_d(y)}\big[-\log\big(D(y)\big)\big] + \\ & \mathbb{E}_{x\sim p_d(x)}\big[-\log\big(1-D\big(G(x,g)\big)\big)\big]. \end{aligned} \tag{11}$$

3.4 Algorithm

In the training phase, the input is the occluded image x_i and the output is the de-occluded image $\hat{y}_i$. The values of all images are scaled to $[-1, 1]$. The details of the training phase are summarized in Algorithm 1.

In the testing phase, D is discarded and we use the trained V and G with an occluded face image as input to get the de-occluded face image. We first use V to get the predicted overall structure g_i of the input occluded image x_i. Then, x_i and g_i are concatenated together and input into G to get the final de-occluded image $\hat{y}_i$.

Algorithm 1 Training phase of the proposed framework.

Input: Occluded face images $\{x_i\} \in X$;
Corresponding ground truth images $\{y_i\} \in Y$.
Output: De-occluded face images $\{\hat{y}_i\}$.

while V not converged **do**
  Sample batch images $\{x_i\}$ from X;
  Predict $\{g_i\}$ for $\{x_i\}$;
  Update V with Eq. (4);
end while
while G not converged **do**
  Sample batch images $\{x_i\}$ from X;
  Use trained V to predict $\{g_i\}$ for $\{x_i\}$;
  Generate de-occluded face images $\{\hat{y}_i\}$ with $\{x_i\}$ and $\{g_i\}$;
  Update G with Eq. (8);
  Update D with Eq. (11);
end while

4 EXPERIMENT

4.1 Datasets and Implementation Details

We synthesize the occluded face image dataset on benchmark datasets, i.e. CelebA and LFW. First, 15 commonly seen occlusion objects can be obtained manually or via a color image segmentation

method, e.g. see [23] and [7]. Then, we overlay these occlusions onto the non-occluded face image with random shape, location, rotation and size. In addition, we preprocess the images by cropping out the face part and then re-scaling them into the size of $128 \times 128 \times 3$ pixels. Figure 5 shows a snapshot of the occluded faces and Figure 6 shows examples of the occlusion objects.

CelebA is a public medium scale face attributes dataset which contains 202,599 celebrity images with 10,177 identities. LFW is a commonly used testing set and contains 13,233 faces of 5,749 individuals collected in uncontrolled environments. The identities in the training phase are separated from the testing phase. To train our proposed FSG-GAN, we randomly select 90% identities in CelebA. The remaining 10% identities in CelebA and the whole LFW are used as testing set. Four testing sets from CelebA and LFW datasets are prepared and the composition details of the training and testing sets are listed in Table 1. The new identities/occlusions mean that the identities/occlusions have not been seen by the model during the training phase. The hyper-parameters in the generator are empirically set at: $\lambda_{pw} = 100$, $\lambda_{ip} = 50$ and $\lambda_{adv} = 1$.

Table 1: Dataset composition details.

Dateset name	New identities	New occlusions
	Training set	
Set-CelebA	×	×
	Testing set	
Set1-CelebA	✓	×
Set2-CelebA	✓	✓
Set1-LFW	✓	×
Set2-LFW	✓	✓

Figure 5: Examples of the synthesized dataset, where images in the top row are from CelebA and those in the bottom row are from LFW.

Figure 6: Examples of the occlusion objects.

4.2 Performance Evaluation

The qualitative index is evaluated by the visual effect of the de-occlusion results. To quantitatively evaluate the de-occlusion performance, two commonly used metrics, namely peak signal-to-noise ratio (PSNR) and structural similarity (SSIM), are used. It is worth noting that face de-occlusion aims at generating identity-preserved occlusion-free face images rather than the same pixels in the original images. Therefore, recognition accuracy is also used to help evaluate the performance of the proposed model.

We compare the proposed FSG-GAN with three state-of-the-art methods, namely Cycle-GAN [51], DeepFill [49] and OA-GAN [12]. Cycle-GAN is a style transform method that can transform the image from the occluded face image domain to the occlusion-free image domain. DeepFill is an inpainting method that shows potential application in removing occlusions. OA-GAN is also a two-stage face image de-occlusion method that performs well in removing facial occlusions. For Cycle-GAN, we retrain the official code with our dataset for fair comparisons. Since DeepFill is also trained on the CelebA dataset, we use the officially provided code and model parameters. As the code of OA-GAN is not publicly available, we therefore re-implemented it by ourselves.

4.2.1 Qualitative Comparison. Figure 7 shows the results of de-occluded face images on CelebA and LFW. Cycle-GAN can perform holistic style transformation and generate semantically meaningful pixels. However, the local detail and identity information are without guarantee of preservation after transformation. DeepFill aims to fill in irregular shaped holes in an image. It generates the detailed pixels but without the preservation of identity, which is therefore essentially inapplicable to practical applications like identification and recognition. Specifically, we can see that in Set1-LFW, DeepFill increases PSNR, but decreases SSIM. That means that DeepFill cannot reconstruct the overall structure of face image well when the occlusion breaks the image structure. In addition, it needs the binary mask of the occlusion object during testing. The reason why OA-GAN does not work well is that the method can only detect and remove small occlusions. The method loses overall consistency because there is no reference information when the occluded regions possess independent semantics (e.g., eyes, mouth and forehead). The facial structure predicted by VAE in the proposed method serves as background knowledge that can supervise the generation. The generator has a large receptive field without reducing the size of the feature maps, which can maintain the overall structure and color consistency. The results show that FSG-GAN successfully removes the occlusion and recovers a photorealistic face image.

4.2.2 Quantitative Comparison. Table 2 shows the comparison in terms of PSNR and SSIM. However, PSNR and SSIM favor the images which are exactly the same as the ground truth. The recognition accuracy of the occluded images is also reported to help evaluate the effectiveness of our proposed FSG-GAN. We choose 200 identities who have more than 6 images from Set1-CelebA and Set1-lfw to calculate the recognition accuracy. In addition, these occluded images simulate natural occlusion. Figure 8 shows a snapshot of these images. For each person, we choose one image as the gallery image while the other five images are used as query images. We repeat the experiment five times and report the average recognition

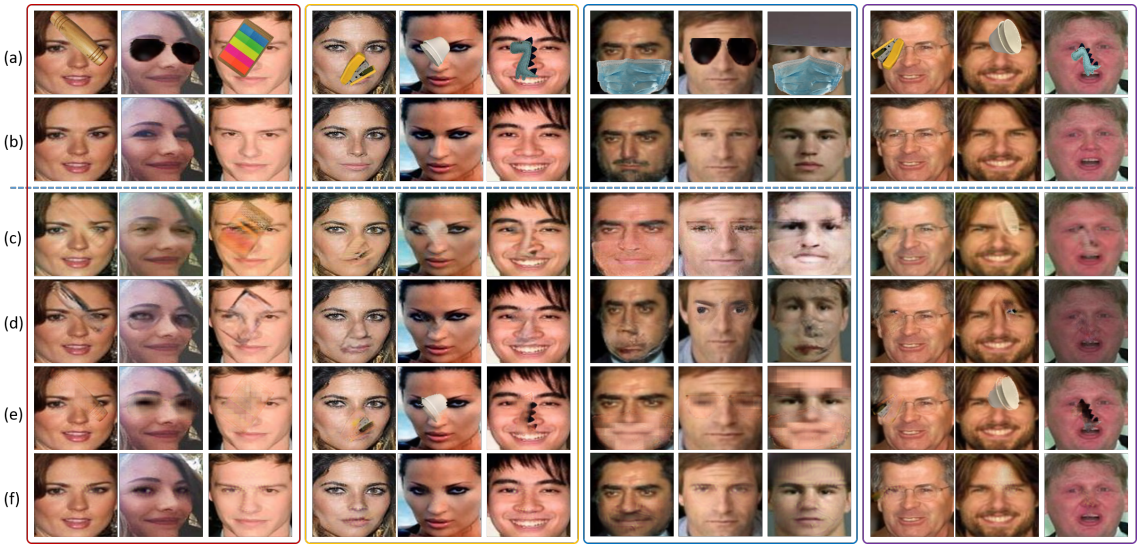

Figure 7: Examples of qualitative comparison, where (a) and (b) are the input occluded face images and the ground truth images, while from (c) to (f) are the de-occlusion results given by Cycle-GAN, DeepFill, OA-GAN, and our proposed FSG-GAN respectively. The boxes with the size of 6×3 images each from left to right show the experimental results on Set1-CelebA, Set2-CelebA, Set1-LFW and Set2-LFW, respectively.

Table 2: Quantitative comparison in terms of PSNR and SSIM.

Method \ Metric	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
	Set1-CelebA		Set2-CelebA		Set1-LFW		Set2-LFW	
Occluded image	10.4764	0.6425	13.6658	0.6852	12.4447	0.7606	14.1494	0.7601
Cycle-GAN (2017)	13.7667	0.6459	15.8253	0.8325	16.0223	0.8238	16.1571	0.8545
DeepFill (2018)	14.5140	0.7029	20.2123	0.8581	14.9931	0.7387	19.4776	0.8591
OA-GAN (2020)	17.1828	0.7215	14.7253	0.7325	19.1809	0.7978	15.0563	0.8458
Proposed FSG-GAN	**21.1112**	**0.7936**	**20.3914**	**0.8699**	**21.9961**	**0.8344**	**19.7848**	**0.8710**

Table 3: Quantitative comparison in terms of recognition accuracy (%).

Method \ Acc.	Rank-1	Rank-5	Rank-10
	Set2-CelebA		
Occluded img	44.2 ± 3.0	54.4 ± 2.5	59.6 ± 2.0
Cycle-GAN	36.9 ± 3.5	51.6 ± 2.8	58.2 ± 2.7
DeepFill	45.1 ± 2.4	59.6 ± 1.4	65.3 ± 0.6
OA-GAN	47.2 ± 2.4	61.9 ± 2.2	68.9 ± 0.9
Proposed FSG-GAN	**64.2 ± 4.7**	**76.3 ± 3.3**	**80.3 ± 2.4**
	Set2-LFW		
Occluded img	52.7 ± 1.6	60.6 ± 1.3	64.0 ± 1.4
Cycle-GAN	46.8 ± 2.6	55.7 ± 1.5	61.4 ± 2.0
DeepFill	52.9 ± 4.4	64.3 ± 2.3	68.2 ± 2.1
OA-GAN	53.1 ± 2.3	64.9 ± 2.1	69.4 ± 2.0
Proposed FSG-GAN	**68.6 ± 4.6**	**77.6 ± 3.4**	**80.1 ± 2.7**

accuracies of the de-occlusion results by these methods on CelebA and LFW. The rank-1, rank-5 and rank-10 recognition accuracies are calculated, respectively, by first extracting deep features with ArcFace [11] and then using a cosine-distance metric to calculate. Table 3 shows the recognition accuracy. The proposed FSG-GAN obtains higher scores in both PSNR and SSIM as well as recognition accuracy, which demonstrates the better performance of FSG-GAN in both image quality and identity preservation.

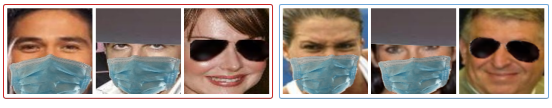

Figure 8: Examples of the images used to calculate recognition accuracy, where (a) is from CelebA and (b) is from LFW.

4.3 Ablation Experiment

To verify the effectiveness of the proposed modules, we conducted two more experiments: train the model (i) without HDC modules and (ii) with regular discriminator that is the same as DeepFill and

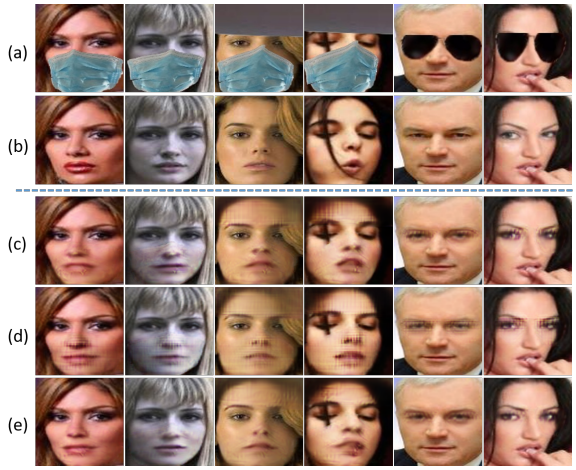

Figure 9: The effectiveness of different modules, where (a) are the input occluded face images and (b) are the ground truth images, while from (c) to (e) are the de-occluded images by the network w/o HDC module, with Regular discriminator, and with HDC module and multi-RFs discriminator respectively.

Table 4: Quantitative evaluation of different modules.

Method \ Metric	PSNR	SSIM
Occluded img	7.9484	0.5298
w/o HDC	17.7813	0.6856
Regular Dis.	17.3046	0.6581
Hdc + multi-RFs Dis.	**18.1103**	**0.6979**

OA-GAN. We utilized the same images that were utilized to calculate the recognition accuracy to demonstrate the effect. Figure 9 shows the qualitative effect and Table 4 reflects the quantitative effect. From row c of Figure 9, it can be seen that the recovered part of the skin tone has less consistency with the original region to a certain extent when the network is without HDC modules. There are unnatural textures in the recovered part with the regular discriminator. Since the regular discriminator only determines the overall similarity without paying attention to image details, it cannot provide motivation for the generator to output the detailed images well.

5 CONCLUDING REMARKS

In this work, we have proposed the FSG-GAN network to address the challenging problem of face image de-occlusion. We first use the VAE to predict the overall facial structure to get the human face topological structure background knowledge. Then, the predicted structure and the occluded image are fed into the following novel U-Net structure to remove the occlusion. Specifically, we have introduced the HDC modules to improve the holistic consistency of the face image. The de-occlusion results are further enforced through the proposed multi-RFs discriminator by distinguishing the different level features of real and generated images. Experiments have shown that the proposed method outperforms the existing counterparts.

Figure 10: Examples of incapable situations, where (a) shows input occluded face images, (b) shows ground truth images, and (c) shows unsatisfactory de-occlusion results.

Even though our model is able to remove various occlusions and generate semantically reasonable and visually realistic contents, it is incapable of handling some situations. One situation is the excessive range of posture. We have used the roughly aligned images in CelebA dataset and implemented various data augmentation to improve the robustness towards different posture, but found that the occlusion in face image with large posture cannot be removed well. We show the examples of the results in the first and second columns of Figure 10. These unsatisfactory de-occlusion results indicate that the model cannot distinguish facial structure semantic information well. Nevertheless, we can exploit contextual attention to alleviate this issue. The other powerless situation is that the occlusion is too severe. When occlusions cover almost all the facial information, the network fails to predict the correct facial structure, leading to unsatisfactory results. The third to fifth columns of Figure 10 demonstrate this situation. That is, the occluded area and the occlusion type have a significant impact on recovery results. We will conduct a systematic research towards these in our future work.

ACKNOWLEDGMENT

This work was supported by ITF grant: ITS/339/18 and HKBU grant: RC-FNRA-IG/18-19/SCI/03.

REFERENCES

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B. Goldman. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics* 28, 3 (2009), 1–12. https://doi.org/10.1145/1531326.1531330

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. 2000. Image Inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. ACM Press/Addison-Wesley Publishing Co., USA, 417–424. https://doi.org/10.1145/344779.344972

[3] Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., USA, 187–194.

[4] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. 2016. Importance Weighted Autoencoders. arXiv:1509.00519 [cs.LG]

[5] Jiancheng Cai, Han Hu, Shiguang Shan, and Xilin Chen. 2019. Fcsr-gan: End-to-end learning for joint face completion and super-resolution. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–8. https://doi.org/10.1109/FG.2019.8756607

[6] Tony F. Chan and Jianhong Shen. 2001. Nontexture inpainting by curvature-driven diffusions. *Journal of Visual Communication and Image Representation* 12, 4 (2001), 436–449. https://doi.org/10.1006/jvci.2001.0487

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

[9] Yiu-ming Cheung, Xin Liu, and Xinge You. 2012. A local region based approach to lip tracking. *Pattern Recognition* 45, 9 (2012), 3336–3347.

[10] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. 2012. Image melding: Combining inconsistent images using patch-based synthesis. *ACM Transactions on Graphics* 31, 4 (2012). https://doi.org/10.1145/2185520.2185578

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanosl Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4690–4699.

[12] Jiayuan Dong, Liyan Zhang, Hanwang Zhang, and Weichen Liu. 2020. Occlusion-Aware GAN for Face De-Occlusion in the Wild. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6. https://doi.org/10.1109/ICME46284.2020.9102788

[13] Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. 2017. Generative Multi-Adversarial Networks. arXiv:1611.01673 [cs.LG]

[14] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. 2018. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision* 126, 12 (2018), 1269–1287. https://doi.org/10.1007/s11263-018-1064-8

[15] Selim Esedoglu and Jianhong Shen. 2002. Digital inpainting based on the Mumford-Shah-Euler image model. *European Journal of Applied Mathematics* 13, 4 (2002), 353–370. https://doi.org/10.1017/S0956792502004904

[16] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. 2015. Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning (ICML)*. 1462—-1471.

[17] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. 2015. Effective Face Frontalization in Unconstrained Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* Technical Report 07-49. University of Massachusetts, Amherst.

[19] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. 2017. 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 (CVPR)*. 2439–2448.

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-To-Image Translation With Conditional Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1125–1134.

[21] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2017. Neural Machine Translation in Linear Time. arXiv:1610.10099 [cs.CL]

[22] Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[23] Lap-tak Law and Yiu-ming Cheung. 2003. Color image segmentation using rival penalized controlled competitive learning. In *Proceedings of the International Joint Conference on Neural Networks*, Vol. 1. 108–112. https://doi.org/10.1109/IJCNN.2003.1223306

[24] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 4681–4690.

[25] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. 2017. Generative Face Completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[26] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. 2019. Face Parsing With RoI Tanh-Warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5654–5663.

[27] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. 2019. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*. 4170–4179.

[28] Xin Liu, Yiu-ming Cheung, Meng Li, and Hailin Liu. 2010. A lip contour extraction method using localized active contour model with automatic parameter selection. In *2010 20th International Conference on Pattern Recognition*. IEEE, 4332–4335. https://doi.org/10.1109/ICPR.2010.1053

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *The IEEE International Conference on Computer Vision*. IEEE, 3730–3738.

[30] Florian Luisier, Thierry Blu, and Michael Unser. 2007. A new SURE approach to image denoising: Interscale orthonormal wavelet thresholding. *IEEE Transactions on image processing* 16, 3 (2007), 593–606. https://doi.org/10.1109/TIP.2007.891064

[31] Albert Michotte, Georges Thinès, and Geneviève Crabbé. 1991. Amodal completion of perceptual structures. *Michotte's experimental phenomenology of perception* (1991), 140–167.

[32] Bence Nanay. 2007. Four theories of amodal perception. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 29. CogSci, USA, 1331–1336.

[33] Bruno A Olshausen and David J Field. 1997. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research* 37, 23 (1997), 3311–3325. https://doi.org/10.1016/S0042-6989(97)00169-7

[34] Meng Pang, Yiu-Ming Cheung, Binghui Wang, and Jian Lou. 2019. Synergistic Generic Learning for Face Recognition From a Contaminated Single Sample per Person. *IEEE Transactions on Information Forensics and Security* 15 (2019), 195–209. https://doi.org/10.1109/TIFS.2019.2919950

[35] Javier Portilla, Vasily Strela, Martin J Wainwright, and Eero P Simoncelli. 2003. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions on Image processing* 12, 11 (2003), 1338–1351. https://doi.org/10.1109/TIP.2003.818640

[36] Rajesh Ranganath, Dustin Tran, and David Blei. 2016. Hierarchical variational models. In *International Conference on Machine Learning (ICML)*. 324–333.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (Eds.). Springer International Publishing, Cham, 234–241.

[38] Christos Sagonas, Yannis Panagakis, Stefanos Zafeiriou, and Maja Pantic. 2015. Robust statistical face frontalization. In *Proceedings of the IEEE international conference on computer vision (CVPR)*. 3871–3879.

[39] Jianhong Shen and Tony F Chan. 2002. Mathematical Models for Local Nontexture Inpaintings. *SIAM J. Appl. Math.* 62, 3 (January 2002), 1019–1043. https://doi.org/10.1137/S0036139900368844

[40] Casper Kaae Sø nderby, Tapani Raiko, Lars Maalø e, Søren Kaae Sø nderby, and Ole Winther. 2016. Ladder Variational Autoencoders. In *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., 3738–3746.

[41] Arash Vahdat and Jan Kautz. 2021. NVAE: A Deep Hierarchical Variational Autoencoder. arXiv:2007.03898 [stat.ML]

[42] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. arXiv:1609.03499 [cs.SD]

[43] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. 2018. Additive margin softmax for face verification. *IEEE Signal Processing Letters* 25, 7 (2018), 926–930. https://doi.org/10.1109/LSP.2018.2822810

[44] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. 2017. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*. 1041–1049. https://doi.org/10.1145/3123266.3123359

[45] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. 2018. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1451–1460. https://doi.org/10.1109/WACV.2018.00163

[46] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. 2008. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence* 31, 2 (2008), 210–227. https://doi.org/10.1109/TPAMI.2008.79

[47] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. 2018. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security* 13, 11 (2018), 2884–2896. https://doi.org/10.1109/TIFS.2018.2833032

[48] Raymond Yeh, Chen Chen, Teck Yian Lim, Mark Hasegawa-Johnson, and Minh N Do. 2016. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539* 2, 3 (2016).

[49] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2018. Generative Image Inpainting With Contextual Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5505–5514.

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

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