FD-StackGAN: Face De-occlusion Using Stacked Generative Adversarial Networks

Abdul Jabbar[1], Xi Li[1*], M. Munawwar Iqbal[2], and Arif Jamal Malik[3]
[1] College of Computer Science, Zhejiang University,
Hangzhou, China
[e-mail: jabbar@zju.edu.cn, xilizju@zju.edu.cn]
[2] Institute of Information Technology, Quaid-i-Azam University,
Islamabad, Pakistan
[e-mail: mmic@qau.edu.pk]
[3] Department of Software Engineering, Foundation University,
Islamabad, Pakistan
[e-mail: arif.malik@fui.edu.pk]
[*]Corresponding author: Xi Li

Received January 23, 2021; revised April 15, 2021; revised May 30, 2021; accepted June 20, 2021; published July 31, 2021

Abstract

It has been widely acknowledged that occlusion impairments adversely distress many face recognition algorithms' performance. Therefore, it is crucial to solving the problem of face image occlusion in face recognition. To solve the image occlusion problem in face recognition, this paper aims to automatically de-occlude the human face majority or discriminative regions to improve face recognition performance. To achieve this, we decompose the generative process into two key stages and employ a separate generative adversarial network (GAN)-based network in both stages. The first stage generates an initial coarse face image without an occlusion mask. The second stage refines the result from the first stage by forcing it closer to real face images or ground truth. To increase the performance and minimize the artifacts in the generated result, a new refine loss (e.g., reconstruction loss, perceptual loss, and adversarial loss) is used to determine all differences between the generated de-occluded face image and ground truth. Furthermore, we build occluded face images and corresponding occlusion-free face images dataset. We trained our model on this new dataset and later tested it on real-world face images. The experiment results (qualitative and quantitative) and the comparative study confirm the robustness and effectiveness of the proposed work in removing challenging occlusion masks with various structures, sizes, shapes, types, and positions.

Keywords: Generative adversarial network (GAN), image restoration, image reconstruction, occlusions mask removal.

http://doi.org/10.3837/tiis.2021.07.014 **ISSN : 1976-7277**

1. Introduction

Revealing human face identity, often corrupted by serious occlusion masks, is one of the most vigorous and widespread study hotspots in computer vision applications, including law enforcement and entertainment systems. Clear face images play the most substantial role in describing face identity characteristics. However, in actual situations in the special events celebration, surveillance cameras or face recognition systems encounter new challenges in which they become inapplicable due to severe occlusion by occlusion mask. Removing the occlusion mask covering the human face's discriminative region and correctly restoring the face missing contents might help face recognition. Furthermore, occlusion-free face images can significantly boost human face recognition systems' efficiency and accuracy when only occluded face images of criminal suspects are in access.

A significant improvement has been made in developing image synthesis methods for the last few years, from an occluded face image to an occlusion-free face image transformation task. They produced plausible results; however, they have some un-ignorable defects associated with the affected regions, such as lack of high-frequency information and lack of perceptual information in situations where they have to deal with occlusion masks that have large variations in structures, sizes, shapes, types, and positions in the face image. This is primarily because these methods are trained where occlusion masks, including medical masks, sunglasses, eyeglasses, microphones, scarves, cups, and hands, have less structures, sizes, shapes, types, positions variations in the face image. Unfortunately, their algorithms also show severe deformations and aliasing flaws in their results, especially regions around the eyes. Such degraded results severely affect many computer vision systems, such as recognition, identification, tracking, detection, and classification.

This work aims to improve the performance of computer vision algorithms for face identification/recognition purposes. For this, we present a GAN [1] based network that automatically eliminates the occlusion mask and creates sharp fillings under the affected region. As a result, the completed face looks realistic and natural and steadies the rest of the surrounding area. In the proposed model, Face De-occlusion using Stacked Generative Adversarial Network (FD-StackGAN), a divide-and-conquer scheme is used to divide the mask de-occlusion process into two key stages. The first stage network generates an initial occlusion-free face image with content details as realistic as possible. The second stage network further polishes the initial occlusion-free face image by adding more photorealistic details to make it more visually pleasing and similar to the target image.

An example of face image de-occlusion is shown in **Fig. 1**. By following the well-known "coarse-to-fine structure recovery method," The Stage-I network removes the occlusion mask from the face image. It generates the initial de-occluded face image, which may have undesired artifacts (e.g., blurriness or glaring errors) in the recovered areas. The Stage-II network further polishes the initial de-occluded face image by removing the undesired artifacts or some deficiencies in the Stage-I result and generates the final de-occluded face image by adding more compelling details.

Moreover, we trained the proposed model on a synthetically created dataset and assessed real-world images collected from the Internet. We compared the performance of the proposed model with previous approaches. Several experimental results prove that the proposed model does comparatively well than the other methods.

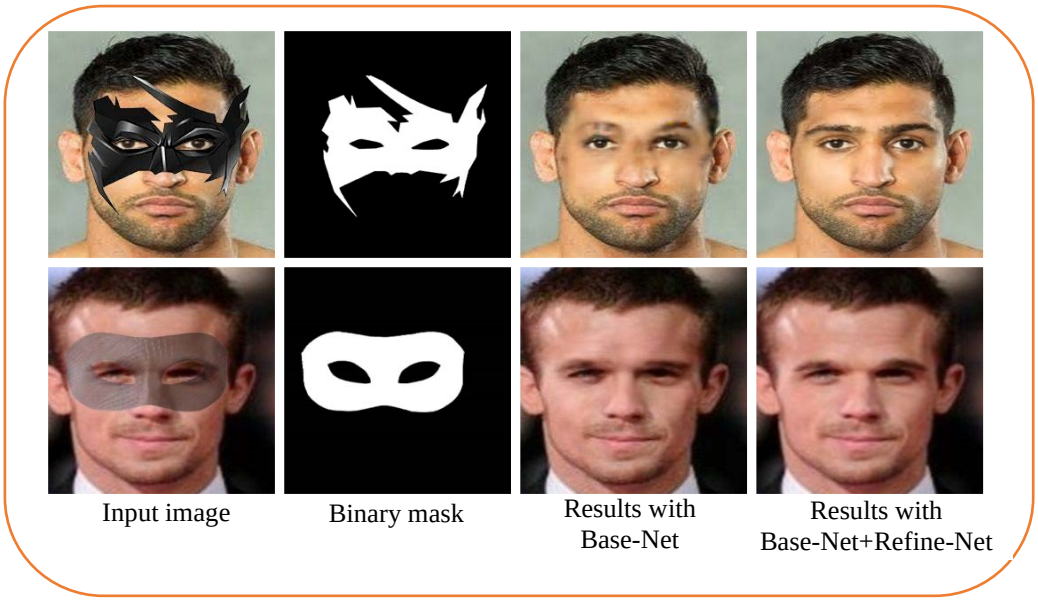

Fig. 1. Results of our model on real-world face images with real occlusion.

The basic structure of this work is as. Related works are presented in Section 2. The proposed approach, along with the loss function, explains in Sections 3. The proposed scheme's implementation and training details are discussed in Section 4. Comparison and discussion, along with ablation studies, are argued in Section 5. Finally, Section 6 concludes the work.

2. Related Work

This segment will discuss various related approaches, classified into traditional methods, CNN-based methods, and GAN-based methods.

2.1 Traditional Methods

Traditional methods can be further specialized into diffusion and patch-based methods. Diffusion-based methods [2]-[4] propagate neighboring information into the missing areas. However, for these diffusion-based approaches, reconstruction is limited to locally accessible information, and these approaches struggle to restore complex structures in missing areas. As a result, these methods cannot effectively handle large absent areas. In contrast, patch-based methods [5], [6] perform well in situations where similar spatial patterns/patches from either the same image or a set of images are copied and pasted into missing regions of the input image. However, these traditional methods are perfect for simple cases where similar and relevant homogeneous patches are propagated from uncovered areas to fill in the small missing area via an iterative searching mechanism of most similar patches. However, these methods cannot produce high-quality face components for large arbitrarily shaped damaged regions with only low-level information.

2.2 CNN-based Methods

The convolutional neural network (CNN) [7] based methods have greatly facilitated image completion advancements. Pathak et al. [8] proposed a context-encoder network (CENet) for image completion. The CENet restores the large missing areas conditioned on its surroundings

data. However, it creates unwanted artifacts and deficiencies in the recovered areas. Iizuka et al. [9] proposed a globally and locally consistent image completion (GLCM) system to complete the arbitrary size missing region in an image. However, it has remarkable noise and artifacts in the recovered region, particularly when holes seem to be at the margins. Zhang et al. [10] proposed a face completion method called DeMeshNet. The DeMehsNet can successfully enforce face identity preservation through perceptual loss, but it fails to recover a face image's large corrupt area. Li et al. [11] introduced a deep symmetry-consistent network (SymmFCNet) for face completion, which forces face symmetry to enhance global consistency. The CNN-based methods can complete arbitrary resolutions and various shapes covered by training a fully convolutional neural network (FCN). Unfortunately, CNN's receptive area is too limited to borrow information from distant spatial locations efficiently. As a result, CNN methods usually produce distorted effects, boundary shadows, and blurred textures.

2.3 GAN-based Methods

The generative adversarial network (GAN), a robust network used for unsupervised machine learning to build a min-max game between two players, i.e., setting up both the player (networks) with their different objectives. One player is called the generator (*G*). The other is the discriminator (*D*). 1st player (*G*) tries to fool the 2nd player (*D*) by producing very natural-looking images from random latent vector z, and 2nd player (*D*) gets better at in-distinguishing between real and generated data. GAN combined cost function is given as:

$$L_{GAN} = V\,(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[log\,(D(x)] + \mathbb{E}_{z\sim p_z(z)}[log\,(1 - D(G(z)))] \qquad (1)$$

Where Equation (1) shows that there are two loss functions, log (*D* (*x*) for the discriminator, and log (1−*D* (*G* (*z*))) for the generator, and two optimizers for the generator and the discriminator since they are two different networks.

The GAN-based methods have revealed encouraging effects in face recovery jobs. Li et al. [12] introduced a GAN-based face completion method (GFCM) for synthesizing the missing contents of a face image. Concerning other methods, this generative face completion method (GFCM) has an additional global discriminator, ensuring a generated face image's reality and enforces the whole face image's consistency. Although the GFCM could produce semantically acceptable results, it has a few shortcomings, such as the image amalgamation operation required to apply the color coherency near the hole borders. The reconstructed face image has particular artifacts, especially when the covered parts are at the borders of an image. Yeh et al. [13] proposed a semantic image completion method using a CGAN [14] on the known region to generate the best un-corrupted image. This method finds the nearby encoding and fills the absent pixels by seeing both the context discriminator and the damaged image. This method has successfully recovered the covered area and generates the absent content well. In the case of huge missing areas, the effects produced are not reliable.

Liao et al. [15] introduced a novel GAN-based collaborative adversarial learning approach called Collaborative GAN (CollaGAN) for face recovery. This CollaGAN suggests that a collaborative adversarial learning method facilitates the direct learning of face completion for better semantic understanding to yield better face in-painting ultimately. The proposed CollaGAN model aims to inductively progress the face completion task by integrating the additional knowledge implanted in other tasks (e.g., landmark detection and semantic segmentation). Introduced a novel GAN-based method for generative image inpainting with novel contextual attention Yu et al. [16] (CA) layer to copy-paste similar feature patches from

nearby related visible regions to the missing regions. The whole network can be trained end-to-end, but the copy-paste approach produces undesired artifacts in the affected region.

Song et al. [17] proposed a two-stage network to perform a face completion task called Geometry Aware Face Completion (GAFC) model. In the first phase, a face geometry estimator estimates the face geometry of the face image. In the second phase, an encoder-decoder generator completes the face image using the face geometry information. Although the model results are superior over many face completion methods, they also severely suffer high computational costs due to prior knowledge of extracting networks. Nazeri et al. [18] proposed a GAN-based Edge-Connect method to recover the image after removing the unwanted objects. EdgeConnect breaks the problem into two stages: edge generator and image completion. The edge generator hallucinates the missing part's edges, followed by the image completion network, filling the missing parts using hallucinated edges. EdgeConnect has successfully recovered the missing areas and generates better results. However, in the case of huge missing regions, it cannot generate a realistic edge map.

Din et al. [19] proposed a two-stage GAN-based framework to remove the face mask and reconstruct the region covered by the mask. The first stage detects the masks, and the second stage gets the reconstructed face. The experimental results outperformed other existing image editing methods. This approach, however, is a difficult and highly time-consuming process. This method also does not generalize well for numerous types of objects (occluded face objects). In most recent times, Maharjan et al. [20] proposed a two-stage GAN-based image-to-image translation method that exploits the full face semantic segmentation instead of the binary segmentation map of the object. The first network is concerned with image translation from occluded face to complete face segmentation. The second network translates the face segmentation map from the first network into a recovered face image. The problem with this method is that this technique is not quite flexible to handle numerous regions containing completely different structures and surrounding backgrounds, especially regions around the eyes, because completed faces must be well-organized the relationship among facial features including eyebrows, eyes, nose, and mouth.

2.4 Relevance to other Works and Significance

After reviewing various related approaches, EdgeConnect, GCA, and GLCM are the closest methods to our work. EdgeConnect also uses a two-staged adversarial approach in which it generates the guidance information in the first stage and edits the image in the second stage. Unlike EdgeConnect, we generate a binary segmentation map of the non-face object while EdgeConnect generates the edge map of the complete image. Moreover, it uses a GAN setup with one discriminator in both stages while we employ two separate discriminators in both stages with two separate generators, which uses CNN-based encoding-decoding network architecture with Skip-connection, which is used in the generator network to strengthen the predictive ability of the generator and to prevent the gradient vanishing caused by the deep network. The result shows that the image completed by the encoder-decoder with skip-connection is more realistic.

In contrast, GLCM and GCA train both discriminators jointly at the same time along with one generator to learn global consistency and deep missing region with a post-processing step like poisson image blending (GL) while we train both discriminators along with two separate generators and our work does not use any supplementary processing or post-processing step. The GLCM and GCA models have noticeable artifacts and blurry in the generated regions since these models predict the missing regions from only high-level features. Different from GLCM and GCA, the proposed model predicts the missing regions from both low-level and high-level

features (pixel-wise loss ($l1$) for low-level features and Structural Similarity loss (SSIM) for high-level features). These schemes (EdgeConnect, GLCM, and GCA) do not work for our problem because they cannot overcome the complexity of the task and produce artifacts due to arbitrary shape large missing regions.

The main difference of this paper from previous work is summarized as follows:

— Different from earlier models, which often fail to generalize well for numerous types of occluded objects (masks) with different shapes and sizes, the proposed model is quite flexible; it can handle numerous occluded regions containing completely different structures, especially regions around the eyes because completed faces must be well-organized the relationship among facial features including eyebrows, eyes, nose, and mouth.
— Unlike earlier models, which often face unstable training problems, the proposed model uses the two time-scale update rule (TTUR) for training. By using different learning rates for generator and discriminator updates, GAN training becomes faster and more stable. We employ the learning rate of 0.0001 for the generator and 0.0004 for the discriminator because a higher learning rate eases the regularized discriminator's slow learning problem.
— The proposed model also shows superiority over several other models in computational efficiency (time cost (s) per sample). Finally, we demonstrate how the proposed model is different from other methods.
— Unlike earlier two-stage models, the proposed model exploits the Stage-I input along with the Stage-I output again as a Stage- II input, as shown in **Fig. 2**. The advantage of using the stage-I input again at Stage- II is that Stage- II can produce high-quality results considering the boundary consistency of the masked region.

Our main contributions are as follows:

— We propose a novel image-to-image translation approach using GAN for face de-occlusion, called Face De-Occlusion using Stacked GAN (FD-StackGAN). FD-StackGAN model can handle face images under challenging conditions, e.g., severe occlusions with significant variations in the structure, size, shape, type, and position in the face image.
— To improve the performance and minimize the artifacts in generated results, a new refine loss function of reconstruction loss (pixel-wise loss ($l1$) for low-level features and Structural Similarity loss (SSIM) for high-level features), perceptual loss, and adversarial loss are introduced to reconstruct well incorporated and visual-artifact-free facial images. This loss could improve the performance of the proposed model.
— To train the proposed model, we have created a new synthetic face image dataset to solve the data scarcity problem by inserting various occlusion masks in facial images using the publicly accessible CelebA dataset.
— Experimental results demonstrate that, although trained on a synthetic face-occluded dataset, the proposed model effectively removes non-face objects and generates structurally and perceptually plausible facial content in challenging real images. The proposed model also illustrates much better computational efficiency in term of time cost (s) for outputting accurate results.

3. Proposed Method

3.1 Overview

This section introduces the proposed multi-stage model for a face image with an occlusion mask to face without occlusion mask transformation tasks. Firstly, we provide details on the structure of FD-StackGAN. Then, we illustrate the considered loss functions in detail. **Fig. 2** shows the framework of the FD-StackGAN. The proposed model takes an occluded face image as input and tries to produce an occlusion-free face image as realistic as possible. This job is accomplished in a coarse-to-fine way: 1) Base-Net at Stage-I and Refine-Net at Stage-II. Each stage model, i.e., Base-Net and Refine-Net, represent a separate GAN. The generator and discriminator of Base-Net are denoted by G_1 and D_1, respectively. While the generator and discriminator of Refine-Net is denoted by G_2 and D_2, respectively. The notation of I_{gt} represents the ground truth. Details will be presented following.

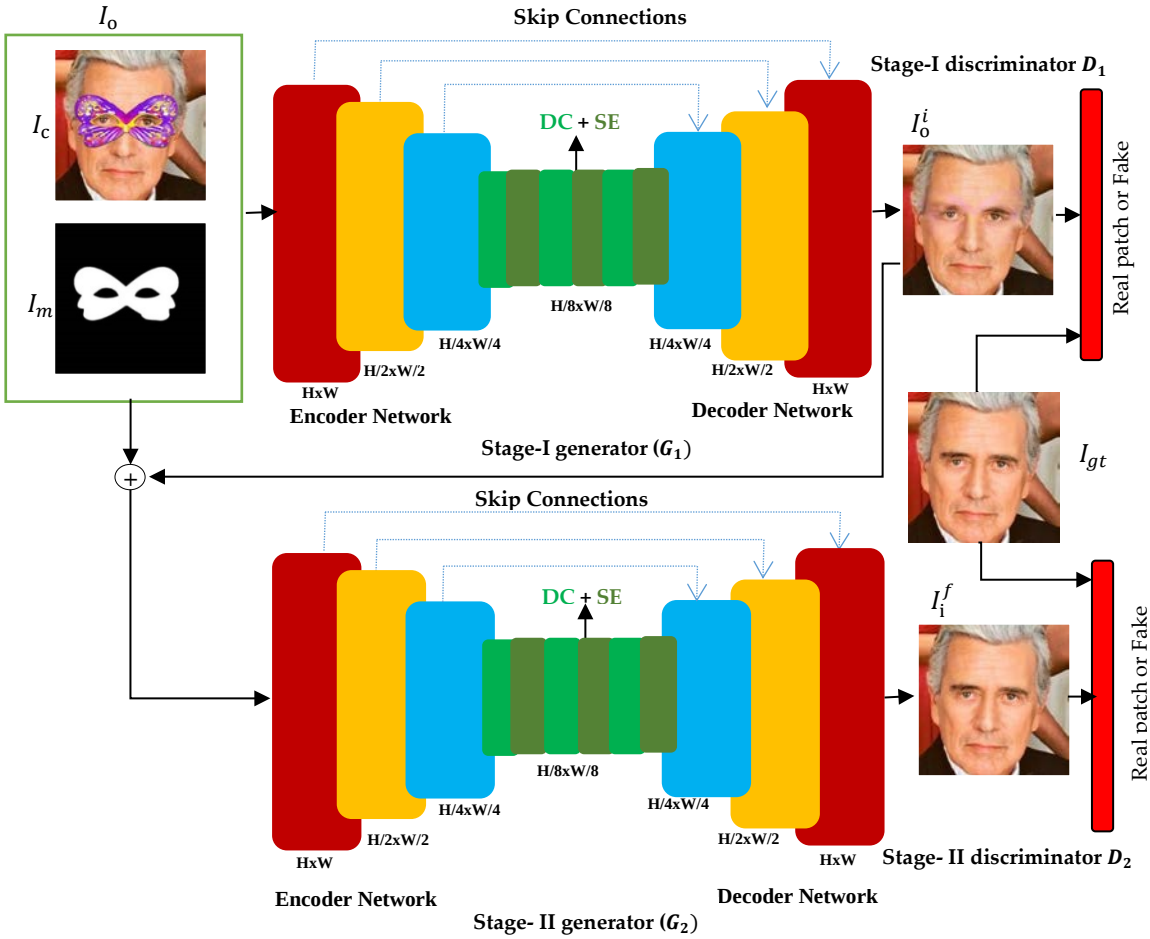

Fig. 2. The proposed FD-StackGAN model architecture. It consists of two separate networks (Base-Net and Refine-Net). The Base-Net generates an initial occlusion-free face image. The Refine-Net refines the Base-Net's generated result by forcing it closer to real face images or ground truth.

3.2 Stage-I: Base-Net

The goal of the Base-Net at Stage-I is to generate the initial coarse occlusion-free face image I_{o}^{i} after taking a face image I_{o} which is a combined input of an occluded face image I_{c} and corresponding binary mask I_{m} of the occluded region. Stage-I generates an initial occlusion-free face image, which is coarse but close to the ground truth.

The generator G_1 in the Base-Net uses CNN-based encoding-decoding network architecture [21]. This encoder-decoder network uses the idea of U-Net [22] with skip connections to stop the loss of spatial information details at higher resolutions during down-sampling and up-sampling functions of the encoder and decoder network. The encoder takes the face image I_{o} as input, and maps it to a low-dimensional latent representation. The decoder network then map back the low-dimensional latent representation, reconstructs and generates the initial coarse output face image I_{o}^{i}. The encoder of the generator G_1 is composed of five convolution layers (for simplicity, only three layers of the encoder network are shown in **Fig. 2**) progressively down-samples the latent representation. Each convolution layer is used in a relu + a convolution + an instance normalization layer, except the first and last layers, which use a tanh in place of a relu. The decoder of the generator G_1 is similar to the encoder except that de-convolution layers substitute convolution layers. The decoder is composed of de-convolution layers, gradually up-samples the latent representation to image scale. A combination of dilated convolution (DC) [23] and Squeeze-and-Excitation' (SE) blocks [24], as shown in **Fig. 2**, is used in the middle of the encoder-decoder. The purpose of dilated convolution (DC) is to enhance the receptive field size without increasing the computational power and network parameters, making the recovered area under the occlusion mask more consistent with its surroundings. The Squeeze-and-Excitation' (SE) block is an addition to the fully convolutionary network (FCN), which enhances a network's representative power by learning the weights for each feature map channel. The SE-blocks re-calibrate feature maps in the context of the channel.

We used a Patch-GAN discriminator D_1 [25] instead of using the regular GAN discriminators to increase the focus on reconstructing high-frequency content. The Patch-GAN discriminator slides a window size of 32 × 32 pixels over the face image I_{o}^{i} and produce a score that shows whether the patch in the face image is real or generated instead of grading the entire face image I_{o}^{i} to produce a more consistent face image with its surroundings.

3.3 Base-Net Loss Function

To minimize the artifacts and ensure better visual quality, a carefully designed arrangement of reconstruction, perceptual, and adversarial loss is used to produce an occlusion-free face image. The Base-Net loss function $L_{\mathrm{stage-I}}$ is composed of a reconstruction loss L_r, a perceptual loss L_p, and an adversarial loss L_a. The Base-Net loss at Stage-I can be expressed as:

$$L_{\mathrm{stage-I}} = \partial L_{\mathrm{r}} + \beta L_p + L_{\mathrm{a}} \tag{2}$$

3.3.1 Reconstruction Loss

The reconstruction loss L_r force the generator to generate more accurate missing content, calculates the difference between the Stage-I generated occlusion-free face image I_{o}^{i} and ground truth I_{gt}. The reconstruction loss is composed: pixel-wise reconstruction loss L_{l1} and structure-level similarity loss L_{SSIM}. The joint reconstruction loss L_r can be stated as:

$$L_r = L_{l1} + L_{SSIM} \tag{3}$$

The pixel-wise reconstruction loss L_{l1} measure the per-pixel difference between Stage-I generated occlusion-free face image I_o^i and ground truth I_{gt}. We calculate the pixel-wise reconstruction loss via l_1-norm in place of l_2-norm because l_1-norm encourages less blurring and glaring errors than l_2-norm. The pixel-wise reconstruction loss L_{l1} can be defined as:

$$L_{l1} = ||I_o^i - I_{gt}|| \tag{4}$$

The structure-level similarity loss L_{SSIM} [26] measure the structural level difference between generated occlusion-free face image I_o^i and ground truth I_{gt}, can be defined as:

$$L_{SSIM} = 1 - \text{SSIM}\left(I_o^i, I_{gt}\right) \tag{5}$$

3.3.2 Perceptual Loss

The perceptual loss L_p encourage the generator output to have identical feature representation to the ground truth measures the feature-level difference between the feature maps of Stage-I generated occlusion-free face image I_o^i and ground-truth I_{gt}, extracted by a well-trained VGG-19 network [27], can be defined as:

$$L_p = \sum_i ||\varphi_i(I_o^i) - \varphi_i(I_{gt})|| \tag{6}$$

3.3.3 Adversarial Loss

The adversarial loss L_a of the Base-Net make the occlusion-free face image I_o^i as close as possible to the target image I_{gt}, and generated realistic results. It can be defined as:

$$L_a = \min_{G_1}\max_{D_1} \mathbb{E}[\log(D_1(I_o, I_{gt})) + [\log(1 - D_1(I_o, G_1(I_o))] \tag{7}$$

3.4 Stage- II: Refine-Net

The GANs have gained considerable attention due to their outstanding data generation capability. However, as GAN has limited learning capacity, considering the uncertainty of filling the missing pixel with semantically plausible and visually pleasing contents, one GAN model may not restore the more delicate texture details. As a result, the face dynamics of the produced images may not be realistic enough. We further process Stage-I results by additional GAN called Refine-Net to generate more realistic images to curb this problem. The Refine-Net at Stage-II adds the necessary refinements to the image generated in the first stage to improve the visual quality further and condense the blurred texture and the inconsistent boundary with the surrounding area. The Refine-Net at Stage-II proves to help further restore the better-quality details to generate a more precise, smoother, and coherent result, especially for the affected region.

Generator G_2 in the Refine-Net at Stage-II is quite similar to the generator G_1 in the Base-Net. We propose the generator G_2 to brings the initial result (Stage-I result I_o^i) closer to the ground truth by rectifying what is missing or wrong in the initial result. To achieve this, we feed I_o (Stage-I input) again with I_o^i (Stage-I output) as a concatenated input into the

generator G_2, which generates the final result I_o^f with more photorealistic details in the recovered area. We feed I_o (Stage-I input) again to enforce edge consistency at the affected region boundary, further increasing the generated face image's visual quality. The Patch-GAN discriminator D_2 of the Refine-Net shares the same architecture as D_1 in the Base-Net. Discriminator D_2 slides a window size of 32 x 32 pixels over I_o^f and produces a score that shows whether the patch in the face image is real or generated.

3.5 Refine-Net Loss Function

Similar to the objective function in Stage-I, we also incorporate the reconstruction loss L_r, a perceptual loss L_p and an adversarial loss L_a in Stage-II. The Refine-Net loss at Stage-II can be expressed as:

$$L_{\text{stage-II}} = \partial L_{\text{r}} + \beta L_p + L_{\text{a}} \tag{8}$$

The adversarial loss L_{a} of the Refine-Net make the recovered face image I_i^f as close as possible to the target image I_{gt}, and generated realistic results. It can be defined as:

$$L_{\text{a}} = \underset{G_2}{min}\underset{D_2}{max}\,\mathbb{E}[\log\,(D_2(I_i^f,\, I_{\text{gt}}))+[\log(1 - D_2(I_i^f, G_2\,(I_i^f))] \tag{9}$$

3.6 Joint Loss Function

Our joint loss improves visually realistic, sharp, and semantically compatible results can be expressed as follows:

$$L_{joint} = \partial L_{\text{r}} + \beta L_p + L_a \tag{10}$$

Where α and β are used to adjust the effect of reconstruction and perceptual loss, respectively.

4. Experiments

We first explain the implementation and training settings of the proposed model in this section. Afterward, we introduce the baseline models. Then, we define datasets and assessment metrics.

4.1 Implementation Details

The Base-Net takes a 256x256 resolution starting occluded face image and generates an occlusion-free face image of the same resolutions. The Refine-Net takes the Base-Net's output face image as input and generates a more real occlusion-free face image with 256x256 resolutions. The proposed two-stage scheme is implemented using Tensorflow deep learning library [28] and is trained with Nvidia GTX 1080Ti GPU to generate 256x256 resolution images. Adam solver [29] trains both stage models (Base-Net and Refine-Net) alternatively for 1000 epochs with a batch size 10. We trained both the stage models for different ∂ and β values. For Stage-I, we used ∂ =100 and β = 33, and for Stage-II, we used ∂ =10 and β = 3.3.

4.2 Training Details (Stabilizing the Training Process of FD-StackGAN)

Although GANs have achieved some incredible results, stable GAN training is a crucial problem. Thus, to avoid unstable GAN training, we used two time-scale update rule (TTUR)

[30] approaches to stabilize FD-StackGAN. Two time-scale update rule approaches use diverse learning levels for the generator and discriminator networks to reach a Nash equilibrium status [31]. We employ the learning rate of 0.0001 for the generator network and 0.0004 for the discriminator network in both stages. Using different learning rates for generator network and discriminator network updates, GAN training becomes more stable because a higher learning rate eases the regularized discriminator's slow learning problem.

4.3 Comparison with Baseline Methods

We compare the performance of the proposed model with the following baseline approaches to demonstrate qualitative and quantitative effects:

— GLCM [9]: A model can complete random size missing regions via globally and locally reliable information.
— GCA [16]: A model for generative image inpainting with novel contextual attention layer to copy-paste similar feature patches from visible regions to the missing regions.
— EdgeConnect [18]: A GAN-based two-stage model recovers the image based on hallucinated edge information from an edge generator network.

4.4 Dataset

We conducted experiments on synthetically generated face images and the real-world face images datasets. The configurations of the two datasets are presented briefly below.

4.4.1 Synthetic Dataset

Occluded face images and corresponding occlusion-free face images are needed to train the proposed model. We train the proposed on our synthesized dataset due to the difficulty in gathering sufficient occluded face images with their corresponding occlusion-free face images. We have synthesized a new face images dataset of 20,000 samples using the available Large-scale Celeb Faces Attributes Dataset (CelebA) [32]. First, we create a synthetic face image dataset by adding different occlusion masks using Adobe Photoshop 2018. **Fig. 3** shows sample images of the occlusion mask used in this dataset**.** Then we create the binary mask of the corresponding occluded region. Each image in our synthetic dataset has a resolution of 256x256. **Fig. 4** shows some sample images and their corresponding masks from this dataset**.**

4.4.2 Real-world Dataset

To demonstrate the proposed method's effectiveness on real-world face images, we build a dataset of real-world images randomly taken from the Internet. While building this occluded face images database, we took all possible care to guarantee that the occluded face images gathered were diverse in size, shape, structure, and position regarding the occlusion mask. These occluded face images are only used for testing (evaluation) purposes. Note that no ground truth exists for real occluded face images because they are downloaded from the Internet.

4.5 Evaluation Metrics

Although the GAN-based models have attained great success in numerous computer vision applications, it is still difficult to evaluate which method (s) is better than other methods because there is no standard defined function for quantitative evaluation, which hurts the GAN performance. Nevertheless, to quantitatively and objectively analyze the accuracy or

effectiveness of the proposed system, we choose various numerical evaluation metrics such as Structural Similarity (SSIM) [33] that guesses the all-inclusive similarity between the reconstructed and the target face images, Peak Signal-to-Noise Ratio (PSNR) that measures the difference in pixel values between the reconstructed and the target face images, Mean Square Error (MSE) that calculates the average squared difference between the reconstructed and the target face images, Naturalness Image Quality Evaluator (NIQE) [34] that measure the quality of image, and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [35] that calculates naturalness of image. For PSNR and SSIM, higher values indicate superior efficiency, while for MSE, NIQE, and BRISQUE, the smaller, the better.

Fig. 3. Some sample images of the occlusion mask used in the synthetic dataset.

Fig. 4. (a) CelebA image, (b) CelebA image with occlusion mask, and (c) binary mask.

5. Comparison and Discussions

This section analyzes the efficiency of the proposed scheme in qualitative and quantitative results. We compare and discuss the proposed model's performances with the baseline models on real-world face images.

5.1 Qualitative Results

In the absence of a robust and consistent evaluation method, the sample quality is mainly evaluated based on the sample's visual fidelity generated by the GAN-based model. Therefore,

we openly displayed the input occluded face images and the output occluded-free face images in the qualitative experiments' test set. For this, we present the qualitative outcomes of the proposed model and the models presented in Section 4.3 (i.e., GLCM [9], GCA [16], and EdgeConnect [18]). **Fig. 5** shows some exemplar results generated by various models. The first column of **Fig. 5** shows the input face images with occlusion mask, while columns 2, 3, and 4 show the samples generated by GLCM, GCA, and EdgeConnect, respectively, with column 5 show the samples generated by the proposed model.

It can be observed in **Fig. 5** that the results of the model are more realistic and smoother than the other three models generated samples, which show visual discrepancies such as blurriness artifacts and the unpleasant boundaries in the recovered region that also show inconsistency with the surrounding area. The proposed model generates visually pleases realistic results because it tries to synthesize more details in the affected region.

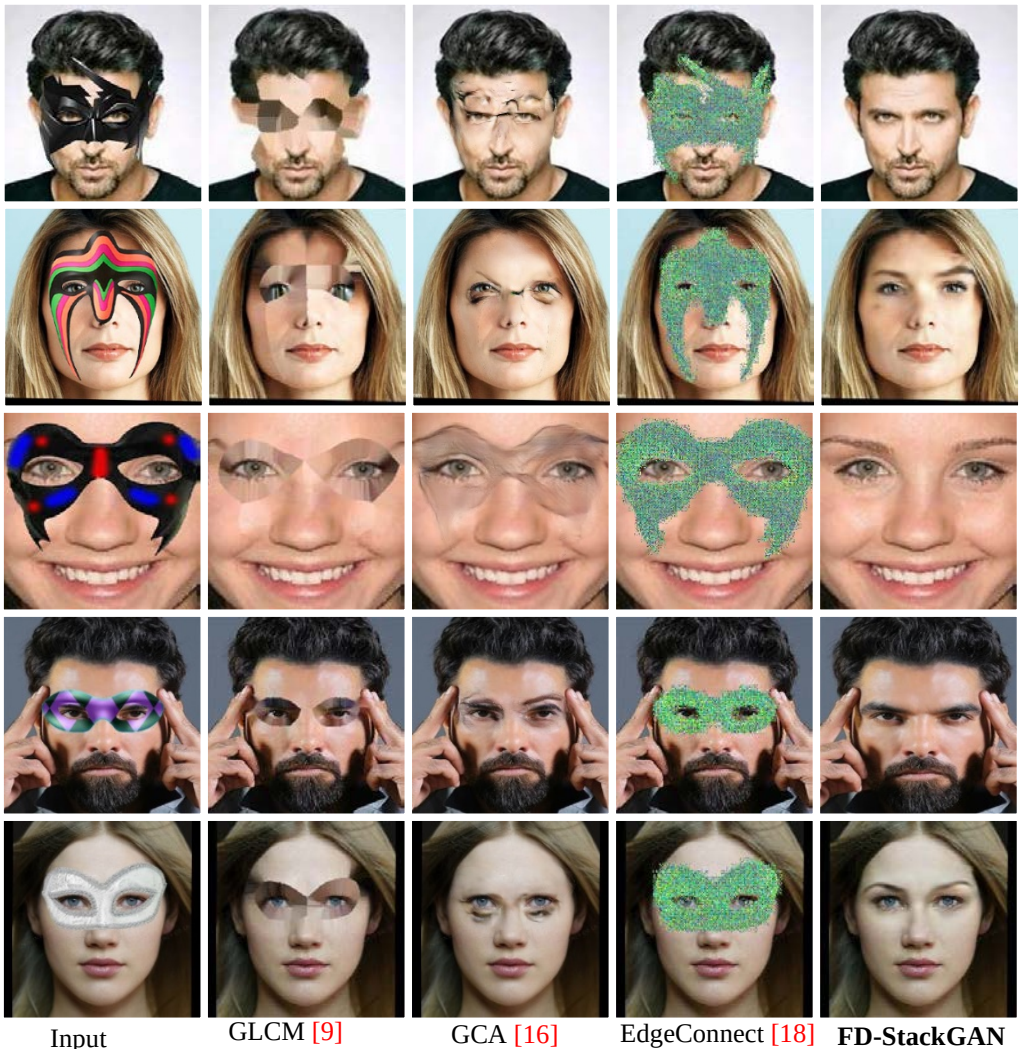

Fig. 5. Visual comparison of our model with baseline models on real-world faces images. As we can see in this figure, the proposed model has better visual results than the other models.

5.2 Quantitative Results

Table 1 shows a quantitative comparison of the proposed model (FD-StackGAN) with baseline representative models (GLCM, GCA, and EdgeConnect). We measure the quantitative performance using five famous evaluation metrics: 1) SSIM, 2) PSNR, 3) MSE, 4) NIQE, and 5) BRISQUE). The quantitative score via SSIM, PSNR, and MSE is calculated using the synthetic test dataset results because no ground-truth exists for real-world occluded face images since they are downloaded from the Internet, while the quantitative score via NIQE and BRISQE is calculated using the results from the real-world test images. The values in **Table 1** were averaged scores obtained from individual test images. From **Table 1**, it has been observed that the proposed model generates semantically consistent and visually plausible face images without occlusion masks, which can help to improve the performance of many computer vision algorithms for face identification/recognition purposes in future studies. The multi-stage proposed approach with careful selection of a new refine loss, e.g., reconstruction loss (pix-wise loss for low-level features, and Structural Similarity loss for high-level features), perceptual loss, and adversarial loss allow in removing the challenging and complex occlusion masks with various structure, size, shape, type, and position. The proposed model shows inferior results compared to other methods on the BRISQUE measurement yet shows superior performance to other assessment metrics.

Please refer to our supplementary material for more quantitative results compared to computational efficiency between the proposed model and baseline models. Finally, we present a quantitative analysis that shows how the proposed model is different from the baseline models. Computational efficiency is an essential practical evaluation metric. It helps researchers monitor the training process and diagnose problems early on or perform early stopping the model during training.

Table 1. Quantitative results comparison of our model with baseline models

Methods	SSIM↑	PSNR↑	MSE↓	NIQE↓	BRISQUE↓
GLCM [9]	0.763	21.953	2329.062	4.754	34.106
GCA [16]	0.797	15.469	2316.839	4.951	**32.761**
EdgeConnect [18]	0.561	15.848	2450.889	16.991	36.426
FD-StackGAN	**0.981**	**32.803**	**34.145**	**4.499**	42.504

5.3 Additional Results

Although we have trained the proposed model using the occlusion masks shown in **Fig. 3** that do not contain cases of occlusion masks used in this experiment. Here, we conducted experiments for some cases of occlusion masks to see how the proposed model works for very different types of occlusion masks from those used in the training occlusion images, as can be seen in the first of rows of **Fig. 6**. As expected, the proposed model produces worse results for real-world face images. The typical failure case behind this is the use of the occlusion masks, which have very different structure types and have very different positions in the face images than the occlusion masks used in the synthetic training dataset that mostly covered the regions around the eyes.

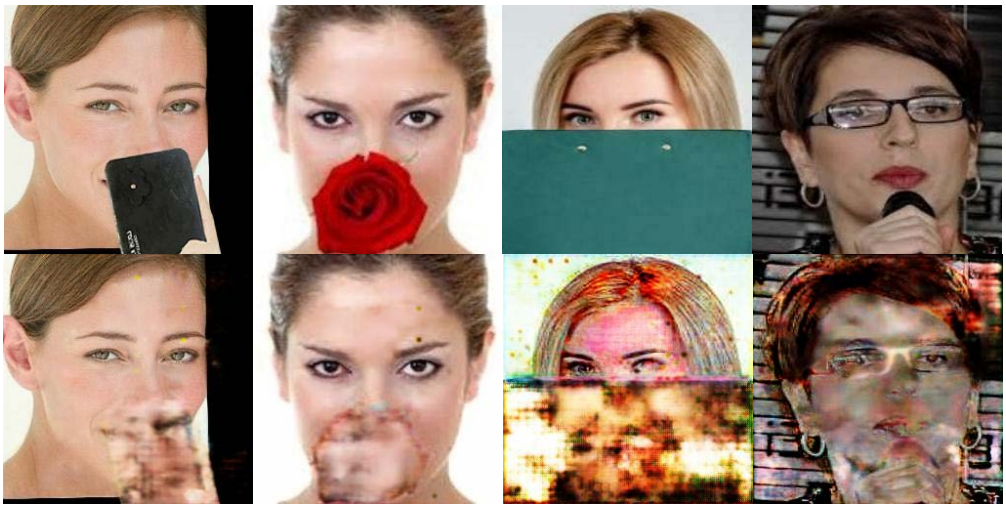

Fig. 6. The performance of the proposed model for real face images with occlusion masks that have very different structure and location in the face images than the occlusion masks used in the synthetic training dataset.

5.4 Ablation Studies

In this subsection, we present the ablation studies to understand the usefulness of using a two-stage network against a single-stage network, and the influence of using a combined loss function.

5.4.1 Comparison between Two Stage Network and Single Stage Network

To demonstrate the effectiveness of using a two-stage network against a single-stage, we perform the ablation study. For this, we make a qualitative and quantitative comparison by training the proposed model with a single-stage network and with a two-stage network. The single-stage model is represented as $G_1 + D_1$ and the two-stage model is represented as G_1+D_1+G_2+D_2. As shown in **Fig. 7**, the proposed model trained with the two-stage can generate more photorealistic results with minimum deformation artifacts than the single-stage results.

The single-stage network generated results are generally blurry with several defects and missing details, especially for occluded regions (red rectangle are used to specify the areas and locations of some undesired artifacts). The two-stage network generated results contain more photorealistic details with minimum undesired artifacts. The two-stage network generates more natural-looking images than a single-stage because the second-stage works as a Refine-Net, i.e., the second-stage corrects what is wrong or missing in the initially recovered regions (blue rectangle are used to specify the areas and locations of some refinement corrections). We also present the quantitative scores of a two-stage network and a single-stage network in SSIM, PSNR, MSE, NIQE, and BRISQUE, as shown in **Table 2**. The numerical scores in **Table 2** clearly showed the advantages of a two-stage network's refinement process over a single-stage model. The two-stage network performed slightly inferior to the single-stage on the SSIM measurement because it is a well-known fact that blurry images often get good SSIM scores despite being less photorealistic and convincing. Yet, the two-stage network is superior to the single-stage on other quantitative measurements.

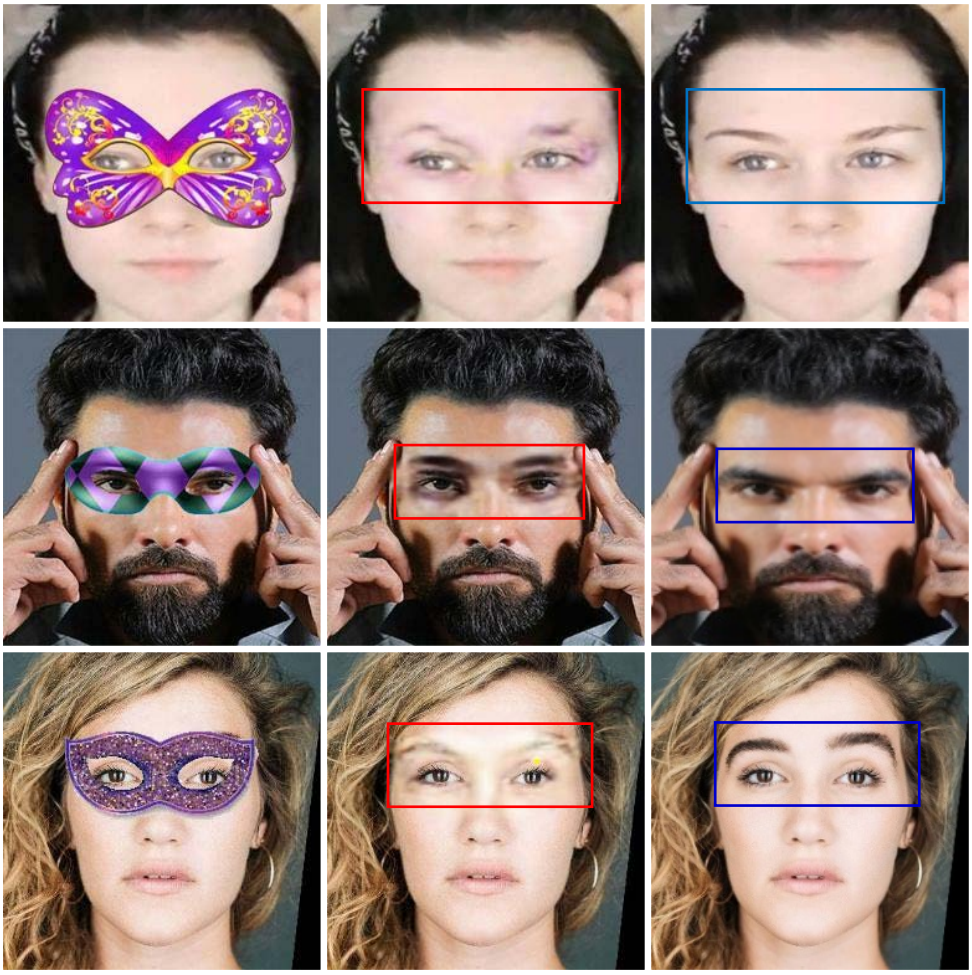

Fig. 7. Effect of using multi-stage networks. From left to right: Input image, results using only Base-Net, and results using both Base-Net + Refine-Net.

Table 2. Quantitative results of a single-stage network and a two-stage network.

Methods	SSIM↑	PSNR↑	MSE↓	NIQE↓	BRISQUE↓
Single-stage network	0.963	30.128	68.860	4.3315	40.2936
Two-stage network	0.981	32.803	34.145	4.3315	40.2936

5.4.2 Integrated Loss Function

Also, we perform the ablation study for joint loss function. For this, we present the ablation studies to analyze the performance of different loss functions. We run the ablation studies to separate the result of pixel-wise reconstruction loss L_{l1} as a reconstruction loss L_r. **Fig. 8** (b) shows the sample results with a reconstruction loss L_{l1}. It seems incapable of recovering the structure of complex vital face semantics, the especially recovered area under the occlusion mask, e.g., area around both eyes. To recover the accurate semantic structure, similarity loss L_{SSIM} is added with a pixel-wise reconstruction loss L_{l1} as a reconstruction loss.

Fig. 8 (c) shows the results with combined reconstruction loss ($L_{l1}+L_{SSIM}$), which successfully recovers the complex vital face semantics of the synthesized face image better than the previous approach but still shows some artifacts. To remove the undesired artifacts and makes the synthesized face image more perceptually closer to the ground truth, we have added the perceptual loss L_P with reconstruction loss (L_{l1}+L_{SSIM}). **Fig. 8** (d) shows the results with joint loss functions ($L_{l1}+L_{SSIM}+L_P$), which tackle the artifacts well and improve the performance. **Table 3** presents the quantitative scores under various loss function settings. Note: This joint loss functions ($L_{l1}+L_{SSIM}+L_P$) shows the Base-Net results only.

Fig. 8. Effects of different loss functions for the Base-Net. From left to right: Input image, results with $\boldsymbol{L_{l1}}$ reconstruction loss only, results with $\boldsymbol{L_{l1}}$+$\boldsymbol{L_{SSIM}}$ joint reconstruction loss, and results with $\boldsymbol{L_{l1}}+\boldsymbol{L_{SSIM}}+\boldsymbol{L_P}$ joint reconstruction loss.

Table 3. Quantitative results under various loss function settings.

Methods	SSIM↑	PSNR↑	MSE↓	NIQE↓	BRISQUE↓
L_{l1}loss	0.984	32.346	43.238	4.7284	41.9486
L_{l1}+L_{SSIM} loss	0.989	32.263	45.501	4.6846	42.1193
$L_{l1}+L_{SSIM}+L_P$ loss	0.963	30.128	68.860	4.3315	40.2936

6. Conclusion and Future Work

This research article proposed an efficient method for face image de-occlusion. The proposed model automatically removes the occlusion mask from the face image and synthesizes the damaged region with compelling details while retaining its original structure. Despite being

trained on a synthetic dataset, its synthesized face images show image quality comparable to other image synthesis methods, which only deal with similar style mask objects. In contrast with earlier methods, the technique here is quite flexible to handle numerous regions containing completely different structures and surrounding backgrounds. In addition, no restrictions are imposed on the topology of the affected region to be inpainted, i.e., the proposed model can successfully remove the numerous types' occlusion masks in the face images by creating semantically useful and visually plausible content for the affected regions. We also analyze the proposed model performance quantitatively and qualitatively and show that the proposed model can produce structurally consistent results of higher perceptual quality. In the future, we plan to extend this work to the videos domain.

Acknowledgment

This work is supported in part by a key scientific-technological innovation research project by the Ministry of Education, Zhejiang Provincial Natural Science Foundation of China under Grant LR19F02000，National Natural Science Foundation of China under Grant U20A20222 and National Key Research and Development Program of China under Grant 2020AAA0107400.

References

[1] I. Goodfellow, J. Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139-144, 2020. Article (CrossRef Link)

[2] S. Esedoglu and J. Shen, "Digital Inpainting Based on the Mumford-Shah-Euler Image Model," *European Journal of Applied Mathematics*, vol. 13, no. 4, pp. 353–370, August 2002. Article (CrossRef Link)

[3] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous Structure and Texture Image Inpainting," *IEEE Transactions* on *Image Processing*, vol. 12, no. 8, pp. 882–889, August 2003. Article (CrossRef Link)

[4] C. Ballester, V. Caselles, and J. Verdera, "Disocclusion by Joint Interpolation of Vector Fields and Gray Levels," *Journal of Multiscale Modeling & Simulation*, vol. 2, no. 1, pp. 80–123, 2004. Article (CrossRef Link)

[5] S. Darabi, E. Shechtman, C. Barnes, Dan B Goldman, and P. Sen, "Image Melding: Combining Inconsistent Images Using Patch-based Synthesis," *ACM Transactions* on *Graphics*, vol. 31, no. 4, pp. 1-10, July 2012. Article (CrossRef Link)

[6] J. Huang and N. Ahuja, "Image Completion using Planar Structure Guidance," *ACM Transactions* on *Graphics*, vol. 33, no. 4, pp. 1–10, July 2014. Article (CrossRef Link)

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "AlexNet," *Adv. of the Neural Information Processing Systems*, vol.1, pp. 1–9, January 2012.

[8] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context Encoders: Feature Learning by Inpainting," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 2536–2544, April 2016. Article (CrossRef Link)

[9] S. Iizuka, E. Simo-serra, and H. Ishikawa, "Globally and Locally Consistent Image Completion," *ACM Transactions on Graphics*, vol. 36, no. 4, pp, 1-14, July 2017. Article (CrossRef Link)

[10] S. Zhang, R. He, Z. Sun, and T. Tan, "DeMeshNet: Blind Face Inpainting for Deep MeshFace Verification," *IEEE Transaction* on *Information Forensics* and *Security*, vol. 13, pp. 637–647, March 2018. Article (CrossRef Link)

[11] X. Li, G. Hu, J. Zhu, W. Zuo, M. Wang, and L. Zhang, "Learning Symmetry Consistent Deep CNNs for Face Completion," *IEEE Transaction on Image Processing*, vol. 29, pp. 7641–7655, December 2018. Article (CrossRef Link)
[12] Y. Li, S. Liu, J. Yang, and M. H. Yang, "Generative Face Completion," in *Proc. of the 30th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5892–5900, January 2017. Article (CrossRef Link)
[13] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic Image Inpainting with Deep Generative Models," in *Proc. of the 30th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6882–6890, January 2017. Article (CrossRef Link)
[14] M. Mirza, and S. Osindero, "Conditional Generative Adversarial Nets," *arXiv preprint arXiv: 1411.1784*, November 2014. Article (CrossRef Link)
[15] H. Liao, G. Funka-Lea, Y. Zheng, J. Luo, and S. Kevin Zhou, "Face Completion with Semantic Knowledge and Collaborative Adversarial Learning," *Lect. Notes Computer Science. (including Subsea. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11361 LNCS, pp. 382–397, 2019. Article (CrossRef Link)
[16] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, "Generative Image Inpainting with Contextual Attention," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. Article (CrossRef Link)
[17] L. Song, J. Cao, L. Song, Y. Hu, and R. He, "Geometry Aware Face Completion and Editing," in *Proc. of the AAAI Conference on Artificial Intelligence*, pp. 2506–2513, July 2019. Article (CrossRef Link)
[18] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning," *arXiv Preprint arXiv:1901.00212*, January 2019. Article (CrossRef Link)
[19] N. U. Din, K. Javed, S. Bae, and J. Yi, "A Novel GAN-Based Network for the Unmasking of Masked Face," *IEEE Access*, vol. 8, pp. 44276–44287, March 2020. Article (CrossRef Link)
[20] R. S. Maharjan, N.U. Din, and J. Yi, "Image-to-Image Translation based Face De-occlusion," in *Proc. of the 12th International Conference on Digital Image Processing*, vol. 11519, June 2020. Article (CrossRef Link)
[21] G. E. Hinton, T. Ms, and R. S. Zemel, "Autoencoders, Minimum Description Length, and Helmholtz Free Energy," *Adv. of the Neural Information Processing System*, pp. 3–10, 2009.
[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 9351, pp. 234–241, November 2015. Article (CrossRef Link)
[23] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution Semantic Image Segmentation," *arXiv preprint arXiv:1706.05587*, June 2017. Article (CrossRef Link)
[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. of the IEEE Conference on Computer Vision Pattern Recognition*, pp. 7132–7141, June 2018. Article (CrossRef Link)
[25] P. Isola, A. A. Efros, B. Ai, and U. C. Berkeley, "Image-to-Image Translation with Conditional Adversarial Networks," *arXiv preprint arXiv:1611.07004v3*, November 2016. Article (CrossRef Link)
[26] 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, April 2004. Article (CrossRef Link)
[27] K. Simonyan, and A. Zisserman, "Very Deep Convolutional Networks for Large Scale Image Recognition," in *Proc. of the International Conference on Learning Representations*, pp. 1–14, April 2015. Article (CrossRef Link)
[28] M. Abadi, P. Barham, J. Chen, et al., "TensorFlow: A System for Large Scale Machine Learning," *Operating System Design and Implementation, Savannah*, vol. 16, pp. 265–283, November 2016. Article (CrossRef Link)

[29] Y. Wang, P. Zhou, and W. Zhong, "An Optimization Strategy Based on Hybrid Algorithm of Adam and SGD," in *Proc. of the 2nd International Conference on Electronic Information Technology*, vol, 232, November 2018. Article (CrossRef Link)
[30] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GAN Trained by a Two Time Scale Update Rule Converge to Local Nash Equilibrium," *Adv. of the Neural Information Processing Systems*, December 2017. Article (CrossRef Link)
[31] L. Ratliff, S. Burden, and S. Sastry, "Characterization and Computation of Local Nash Equilibria in Continuous Games," in *Proc. of Annual Allerton Conference on Communication, Control, and Computing*, pp. 917–924, October 2013. Article (CrossRef Link)
[32] Z. Liu, P. Luo, X. Wang, X. Tang, "Deep Learning Face Attributes in the Wild," in *Proc. of the IEEE International Conference on Computer Vision*, December 2015. Article (CrossRef Link)
[33] 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, April 2004. Article (CrossRef Link)
[34] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, December 2012. Article (CrossRef Link)
[35] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/Referenceless Image Spatial Quality Evaluator," in *Proc. of the 45th Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pp. 723–727, November 2011. Article (CrossRef Link)

Abdul Jabbar is currently a Ph.D. candidate and pursuing a degree at Zhejiang University. His research interests include machine learning, deep learning, generative adversarial networks (GAN), and GAN application to computer vision for image object removal.

Xi Li received the Ph.D. degree from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China, in 2009. From 2009 to 2010, he was a Post-Doctoral Researcher with CNRS Telecom Paris Tech, France. He was a Senior Researcher at the University of Adelaide, Australia. He is currently a Full Professor at Zhejiang University, China. His research interests include visual tracking, motion analysis, face recognition, web data mining and image and video retrieval.

M. Munawwar Iqbal received the Ph.D. degree from the National University of Sciences & Technology, Islamabad, Pakistan, in 2020. He is currently working as an assistant professor in the Institute of Information Technology, Quaid-i-Azam University, Islamabad, Pakistan. His research interests include information fusion, machine learning, and pattern recognition.

Arif Jamal Malik received the Ph.D. degree in Computer Science from the National University of Computer and Emerging Sciences, Islamabad, Pakistan, in 2014. He is currently working as an Assistant Professor at the Department of Software Engineering, Foundation University Islamabad, Pakistan. His research interests include Evolutionary Computation, Swarm Intelligence, Data Science, Machine Learning, and Network Security.