IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 17, 2020, accepted February 25, 2020, date of publication March 2, 2020, date of current version March 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2977386

A Novel GAN-Based Network for Unmasking of Masked Face

NIZAM UD DIN, KAMRAN JAVED, SEHO BAE, AND JUNEHO YI
College of Information and Communication Engineering, Sungkyunkwan University, Suwon 16419, South Korea
Corresponding author: Juneho Yi (jhyi@skku.edu)

ABSTRACT Recent deep learning based image editing methods have achieved promising results for removing object in an image but fail to generate plausible results for removing large objects of complex nature, especially in facial images. The objective of this work is to remove mask objects in facial images. This problem is challenging because (1) most of the time facial masks cover quite a large region of face that even extends beyond the actual face boundary below chin, and (2) facial image pairs with and without mask object do not exist for training. We break the problem into two stages: mask object detection and image completion of the removed mask region. The first stage of our model automatically produces binary segmentation for the mask region. Then, the second stage removes the mask and synthesizes the affected region with fine details while retaining the global coherency of face structure. For this, we have employed a GAN-based network using two discriminators where one discriminator helps learn the global structure of the face and then another discriminator comes in to focus learning on the deep missing region. To train our model in a supervised manner, we create a paired synthetic dataset using publicly available CelebA dataset and evaluated on real world images collected from the Internet. Our model outperforms others representative state-of-the-art approaches both qualitatively and quantitatively.

INDEX TERMS Generative adversarial network, object removal, image editing.

I. INTRODUCTION

The goal of this research, as illustrated in Figure 1, is interaction-free large object (e. g., face mask) removal from facial images. In this work, we focus on unmasking of masked face because it is a very intriguing problem of great practical value. Given an input masked facial image, we detect the mask region, then feed the input image and a binary map of the detected mask region into a GAN [1] based network and generate an image without the non-face object, which is the mask object in our case.

Trend of wearing masks in public is growing in recent years all over the world. First, people wear masks to guard themselves from pollution. Second, some people are self-conscious about their look and they want to hide their face and emotions from the public. Removing the mask object that covers almost half of the face might be of help in guessing one's identity.

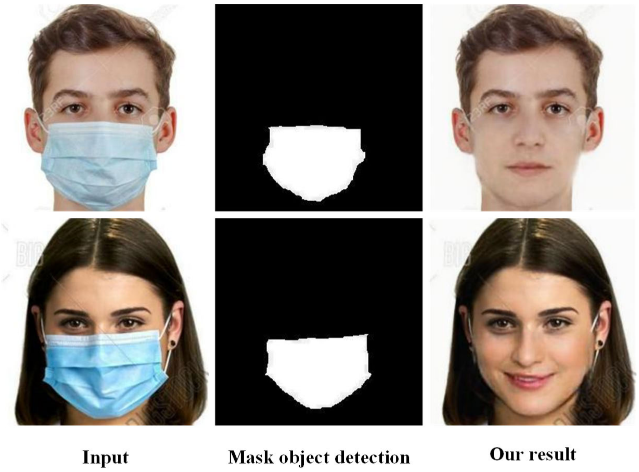


FIGURE 1. Mask removal results of our method on real world images.

To address this task, early non-learning based works [2], [3] erase unwanted object and synthesize the missing content by matching similar patches from the remainder of the image. In [4], they find similar patterns from a database of millions of scene images and paste those patterns in the damaged part. Park *et al.* [5] remove eye glasses from facial images using PCA reconstruction and recursive error compensation. However, these non-learning

The associate editor coordinating the review of this manuscript and approving it for publication was Qiangqiang Yuan.

based algorithms are limited only to small object removal from images.

Recent advances in learning-based methods empower image editing algorithms by learning from large-scale datasets, and thereby outperform non-learning methods for removing unwanted object in an image. Izuka *et al.* [6] use a GAN setup with two discriminators to remove unwanted object and fill damaged region with synthetic content in an image. Two-stage networks have been presented in [7]–[12]. In the first stage, they generate a coarse output and refine it in the second stage. Khan *et al.* [8] also employ a coarse-to-fine network approach to remove microphone object from facial images. SPG-Net [12] and EdgeConnect [11] also use a two-staged adversarial approach to remove unwanted object. In the first stage, they produce some guidance information and complete the image in the second stage using the guidance information from the first stage. We give more detailed descriptions on relevant literature in the related work section.

In general, learning-based image editing approaches work well for removal of objects that have less structural and appearance variations. However, these approaches do not fit well for unmasking of masked face due to large size and complex nature of the object, i. e., mask. For example, most of the time masks cover not only half of the face semantics but also some parts beyond the actual boundary of the face. It starts from upper part of the nose (just below the eyes) and ends up covering some part of the neck, also some parts beyond the cheeks.

To solve this problem, we propose a novel GAN-based network that automatically removes mask and completes the missing hole so that the completed face not only looks natural and realistic but also has consistency with the rest of the image. We break the problem into two: mask object detection and image completion of the detected mask region. In the first stage, we detect the non-face object, i. e., mask, and generate a binary segmentation map of the object using an encoder-decoder network. In the second stage, we take an approach of gradually learning global coherency and deep missing semantics. We first train our model using one generator and one discriminator. This discriminator looks at the whole image and hence help enforcing the global coherency. Although this setup generates the face structure, especially, the chin and cheeks part covered by the mask intact with the rest of the face, but is unable to synthesize well the deep region of the missing hole. By ‘deep region of missing hole’, we mean part of the face far away from the occlusion boundary caused by the mask object, e. g., mouth part of the face, more specifically, lips and teeth. To focus more on generating the deep missing semantics, we add a second discriminator to the model that looks only at the missing region. This scheme enforces the two discriminators to provide fair feedback to the generator to complete the effected region with fine details while maintaining global structure of the facial image. More details of the model training are discussed in the training part of the experiments section. We also introduce a joint loss function that encourages visually plausible, sharp and semantically consistent results.

Moreover, because facial image pairs with and without mask object do not exist, we have created a paired synthetic dataset by editing images from publicly available CelebA dataset.

The main contributions of this work are:

- We propose a novel approach that automatically removes mask object from face and synthesizes the affected region with fine details while retaining the original structure of the face.
- To retain structural and appearance consistency of the recovered face, we use a gradually growing network approach using two discriminators. Where one discriminator first help learning the global structure of face and then another discriminator comes in to focus on learning the deep missing region. This way, we achieve the effect of coarse-to-fine image completion.
- To overcome the data scarcity problem, we have created a synthetic paired dataset using publicly available CelebA dataset.
- Our unified feed forward model generates structurally and perceptually plausible facial image for challenging real images although trained on the synthetic dataset created.

II. RELATED WORK

Object removal from an image consists of two main tasks: a) object detection, b) image completion. There has been a considerable amount of non-learning or learning based work in the field of computer vision to tackle the task of object removal in an image. Due to the plethora of related literature, we only review some representative works related to object detection and completion in an image.

A. OBJECT REMOVAL AND IMAGE COMPLETION

Table 1 shows comparison of our method with non-learning or learning based state-of-the-art object removal approaches. Non-learning based object removal algorithms [2], [4], [13] erase unwanted object from an image and complete the missing region by finding the similar structure from input image or external data. Hays and Efros [4] use thousands of scene images to search information which is most similar to the input sample, and then copy and paste those information into the missing pixels of input sample. In [2], [13], they complete holes left behind of the removed object by extending the surrounding contents into the missing region. However, they produce inconsistent content for images having complex semantic structures and diversified texture, e. g., human faces. Park *et al.* [5] remove eye glasses from facial images by introducing a regularized factor to adjust the patch priority function in computing the filling order. Their work only performs well for removing small object such as eye glasses and fail to generate plausible contents for large objects removal in facial images.

TABLE 1. Comparison of different approaches for object removal.

Methods	Description	Datasets	Goal	Automatic object removal	Deep learning based
A. Criminisi *et al.*, CVPR 2003 [2]	Erase an unwanted portion of an image, and synthesizes a missing region that plausibly matches the remainder of the image	Digital photographs	Object removal	✗	✗
J. Wang *et al.*, Neurocomputing 2014 [13]	Improve [2] by introducing a regularized factor which adjusts the curve of the patch priority function in computing the filling order	Digital photographs	Object removal	✗	✗
J. S. park *et al.*, TPAMI 2005 [5]	Remove eye glasses using PCA reconstruction and recursive error compensation	KFDB (Korean Face Database) [14]	Object (eyeglasses) removal	✓	✗
J. Hays and A. Efros, ACM 2008 [4]	Search similar information from a large database and paste those information into the corrupted pixels of the input sample.	Photographs collected from internet and LabelMe [15]	Object removal	✗	✗
R. Shetty *et al.*, NIPS 2018 [16]	An end-to-end model which learns to find and remove objects automatically from scene images	COCO [17]	Object removal	✓	✓
S. Iizuka *et al.*, ACM Trans. 2017 [6]	Learn global coherency and corrupted region completion jointly using a GAN setup with global and local discriminators	Places2 [18] and CelebA [19]	Image editing	✗	✓
J. Yu *et al.*, CVPR 2018 [9]	Use a coarse-to-fine GAN-based approach with contextual attention module for image inpainting	ImageNet [20], Places2 [18] and CelebA [19]	Image editing	✗	✓
K. Nazeri *et al.*, arXiv preprint 2019 [11]	Use a two-stage adversarial model with edge generator followed by an image editing network for image editing	CelebA [19], Places2 [18] and Paris StreetView [21]	Image editing	✗	✓
M. Khan *et al.*, Electronics 2019 [8]	Used coarse-to-fine GAN-based inpainting approach to remove microphone from facial images	Synthetic dataset created using CelebA [19]	Object (microphone) removal	✗	✓
Ours	Employ a gradually growing GAN setup with two discriminators to remove mask object in facial images	Synthetic dataset created using CelebA [19]	Object removal	✓	✓

On the other hand, learning-based image editing methods outperform those traditional methods both qualitatively and quantitatively. There has been a considerable amount of learning based work on image editing. They mainly describe image inpainting with the main application of object removal. Li *et al.* GFCM [22] and Iizuka *et al.* GLCM [6] train their model to remove an object and reconstruct the damaged part using a GAN setup. To make the generated part locally and globally consistent with rest of the image, GLCM uses two discriminators (global and local discriminator) combined with post processing. Although GLCM completes the image for random damaged region in facial images, it is limited to relatively low resolutions (178 × 218) and produce artifacts when damaged part is at the margins of an image. The output of GFCM also suffers when the removed object is large in size. Dong *et al.* [23] synthesize high-quality results for filling voids of radar data. They use a shadow constrained conditional GAN network to restore the damaged region. However, this work is limited to radar data restoration.

For object removal, Contextual Attention (GCA) [9] and MRGAN [8] use a two stage network. The first stage network produces a coarse result while the second stage network refines the output from the first stage. GCA introduces a contextual attention layer to explicitly attend on related feature patches. MRGAN remove microphone object from facial images. They generate a coarse output for the damaged region only in the first stage and refine it in the second stage. Both GCA and MRGAN generate plausible results for removing small objects but produce unnatural contents for large complex missing region. EdgeConnect [11] and SPG-Net [12] also use a two-staged adversarial approach. Instead of generating a coarse output, they generate an edge map or segmentation map in the first stage. In the second stage, they generate the missing region using the guidance map along with the input image. These schemes do not work for our problem because most of the time the first stage is unable to generate a reasonable map due to the large size of missing region. Moreover, all these deep learning based image editing works [6], [8], [9], [11], [22] assume that users provide the object map at the inference stage. In [16], they automatically detect object region and remove it in general scene-level images. However, their output heavily depends on automatic object detection which oftentimes fails to detect the object region due to large variations in appearance and structure of both mask and face. Moreover, they fill the removed object region by propagating information from surrounding regions. These reasons cause difficulty for the method in [16] to automatically remove large objects from facial images. For object detection we give more reviews at the end of this section.

EdgeConnect [11] is the closest method to our work in a sense that it generates the guidance information in the first stage and edit the image in second stage. Different from EdgeConnect, we generate a binary segmentation map of the non-face object while EdgeConnect generate the edge map of the complete image. Moreover, it uses a GAN setup with one discriminator in both stages while we use a simple encoder-decoder architecture in the first stage for generating binary segmentation map and employ two discriminators in the second stage as in GLCM [6] and GCA [9]. In contrast, GLCM and GCA train both discriminators jointly at the same time along with generator to learn global consistency and deep missing region while we gradually add them to the model.

B. OBJECT DETECTION

R-CNN [24] is a pioneering object detection model that uses deep convolution neural network (CNN) for object detection. It first extracts thousands of regions from an image using selective search algorithm. These regions are then fed into a CNN that produces a feature vector for each proposed region. Finally, SVM classifies the presence of the object within that candidate region proposal from the extracted feature. Fast R-CNN [25] and Faster R-CNN [26] improve the performance of R-CNN by modifying its network architecture. Although these methods produce state-of-the-art results, they require a huge amount of training samples and computation power. Hence, instead of using these expensive algorithms for automatically detecting non-face object in facial images, we employ a simple segmentation network focusing on mask object.

Fully convolutional neural network (FCN) [27] is one of the pioneering end-to-end trained network for image segmentation that use CNN-based auto-encoder setup. FCN encoder is a modified version of popular classification module by replacing the fully connected layers with 1 × 1 convolution. This produces good results though oftentimes fuzzy object boundaries occur. U-Net proposed by Ronneberger *et al.* [28] is one of the most popular end-to-end fully convolutional network in biomedical image segmentation. Encoder captures the context in the image using a series of convolution with max pooling layers while decoder upsamples the encoded information using transposed convolution. Moreover, feature maps from the encoder are concatenated to the feature maps of the decoder. This helps better learning of contextual (relationship between pixels of the image) information. Due to simplicity and better performance of the U-Net architecture, we use it with slight changes to detect the non-face object in the image and generate the corresponding binary segmentation map of the object.

III. APPROACH

The overall structure of our framework is illustrated in Figure 2. It consists of two main modules, map module and editing module. Details of each module are explained in the following.

A. MAP MODULE

The output of the map module is a binary segmentation map, I_{mask_map}, with 1 indicating the mask object and 0 for the remaining pixels in the image. The map generator, G_{mask}, consists of a CNN-based encoder and decoder architecture, which is a modified version of the U-Net [28]. The encoder part of the generator consists of five blocks of convolution layers shown in Figure 2. Here, each block means a convolution layer followed by L*relu* activation function and *instant_norm* layer except the first layer of the encoder. The decoder architecture is a mirror copy of encoder architecture except that convolution is replaced by deconvolution layer. The last layer of the decoder uses *tanh* activation function without normalization layer. Also, we combine local information with the global information by concatenating the result of the deconvolution layers with the feature maps from the encoder at the same level. Usually, these connections are referred as skip connections shown in Figure 2. The map generator network takes an input image I_{input} and is down-sampled to the bottleneck layer using the encoder network. The decoder network is then up-sampled to predict a binary map. We use a cross-entropy loss between the predicted binary map and corresponding target map. To get a clean mask, we take a post processing step by using simple morphological image processing operations of erosion and dilation.

B. EDITING MODULE

The goal of this module is to remove the mask and complete the left behind region in a way that is both structural and appearance wise consistent with the ground truth image. Given the input image I_{input}, guided by the object map I_{mask_map}, our aim is to generate a complete image without the mask. The main blocks of this module are editing generator, discriminators and perceptual network.

1) EDITING GENERATOR

The editing generator, G_{edit}, has the same architecture as the map generator G_{mask}. Different from G_{mask}, we use squeeze and excitation (SE) block [29] at the output of the first three blocks of the encoder. Moreover, between encoder and decoder, we employ four layers of atrous convolution (*rate*: 2,4,8,16) [30], which helps make the missing part generation coherent with rest of the face image by capturing large fields of view. The generator takes the input image, I_{input}, concatenated with the output of the map module, I_{mask_map}, and produce a generated image, I_{edit}.

$$I_{edit} = G_{edit}(I_{input}, I_{mask\_map}). \quad (1)$$

To force the editing generator to produce realistic missing content, we use reconstruction loss which is amalgam of l_1 loss and structural similarity loss *SSIM* [31] , expressed as:

$$\mathcal{L}_{rc} = \mathcal{L}_{l_1} + \mathcal{L}_{ssim}. \quad (2)$$

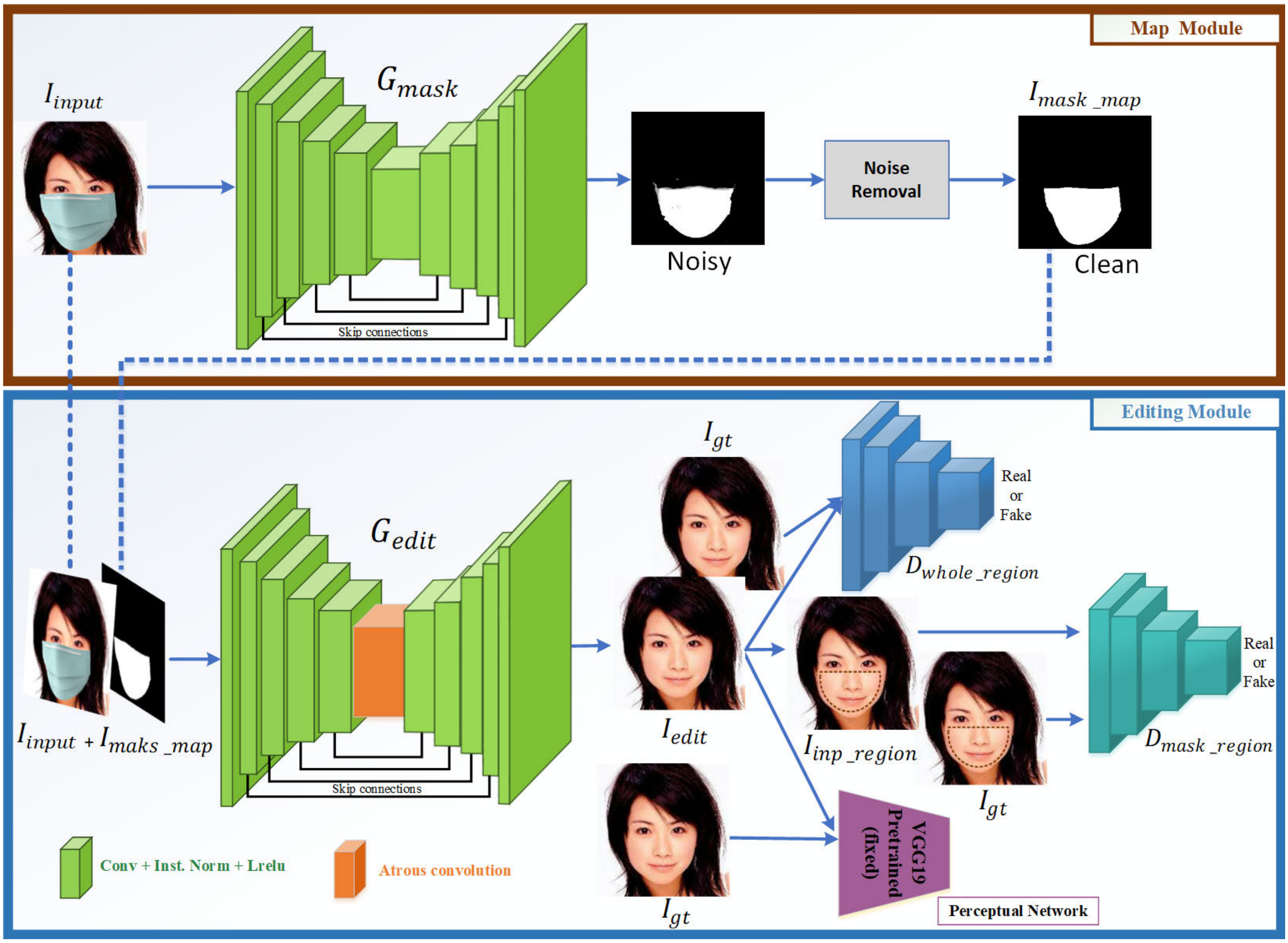


FIGURE 2. **Proposed architecture for our model.**

L_{l_1} loss is the pixel difference between the generated image I_{edit} and the ground truth I_{gt} as:

$$\mathcal{L}_{l_1} = |I_{edit} - I_{gt}|_1. \tag{3}$$

SSIM measures the structural similarity between the I_{edit} and I_{gt} and its corresponding loss function is written as:

$$\mathcal{L}_{ssim} = 1 - SSIM(I_{edit}, I_{gt}). \tag{4}$$

2) DISCRIMINATORS

We use two discriminators called D_{whole_region} and D_{mask_region} as shown in Figure 2. The architecture of both discriminators is the same as the discriminator in pix2pix [32]. They penalizes the dissimilar structure at the patch scale of 70 × 70. The role of both discriminators is to force the editing generator to produce visually plausible and semantically consistent images. Instead of training both discriminators at the same time along with the editing generator, we train the editing generator along with D_{whole_region} for the first 2/5 period of the total training iterations. This helps enforce the output produced by the generator to be structurally consistent with the original input face image by minimizing the following objective function:

$$\begin{aligned}\mathcal{L}_D^{whole\_region} &= -\mathbb{E}_{I_{gt}\in O} \log D_{whole\_region}(I_{edit}, I_{gt}) \\ &\quad + \mathbb{E}_{I_{edit}\in S} \log(1 - D_{whole\_region} \\ &\quad \times (G_{edit}(I_{input}, I_{mask\_map}))) \end{aligned} \tag{5}$$

Here, O and S denote real and synthesized image sets, respectively. However, this loss is not capable of generating plausible content at the deep pixels of the missing region.

We enforce the optimization of D_{mask_region} to produce good semantics in the missing region only. We add D_{mask_region} along with D_{whole_region} to the editing generator. We train them jointly for the rest of the training iterations. To train D_{mask_region}, the following objective function is minimized:

$$\begin{aligned}\mathcal{L}_D^{mask\_region} &= -\mathbb{E}_{I_{gt}\in O} \log D_{mask\_region}(I_{mask\_region}, I_{gt}) \\ &\quad + \mathbb{E}_{I_{edit}\in S} \log(1 - D_{mask\_region}(G_{edit} \\ &\quad \times (I_{input}, I_{mask\_map}))) \end{aligned} \tag{6}$$

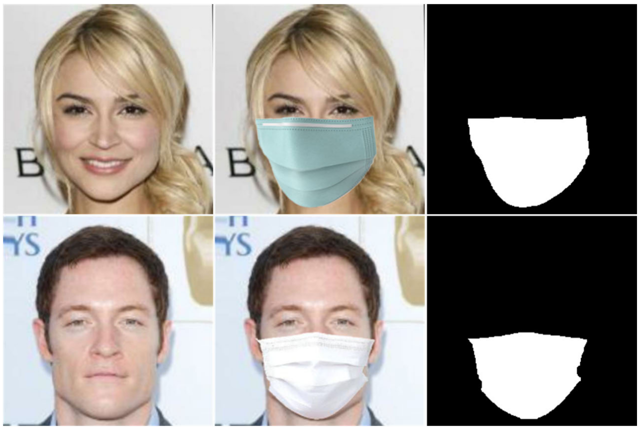


FIGURE 3. Example images of our synthetic training dataset.

Here, $I_{mask\_region} = I_{input} \otimes (1 - I_{mask\_map}) + (I_{edit} \otimes I_{mask\_map})$ and $\otimes$ denotes the element-wise multiplication.

In order to train our model in a GAN setup, the generator fools the discriminators by minimizing the following loss functions:

$$\mathcal{L}_{adv}^{whole\_region} = -\mathbb{E}_{I_{edit} \in S} \log(D_{whole\_region} \times (G_{edit}(I_{input}, I_{mask\_map}))) \quad (7)$$

$$\mathcal{L}_{adv}^{mask\_region} = -\mathbb{E}_{I_{edit} \in S} \log(D_{mask\_region} \times (G_{edit}(I_{input}, I_{mask\_map}))) \quad (8)$$

3) PERCEPTUAL NETWORK

The third block of the editing module is a perceptual network. It is a pre-trained VGG-19 fixed network [33]. The purpose of this network is to encourage the generator output, I_{edit}, to have similar feature representation to the ground truth, I_{gt}. We use a perceptual loss $\mathcal{L}_{perc}$ [34] to penalize the outputs that is perceptually not reasonable by defining a feature level distance measure between the intermediate feature maps of I_{edit} and I_{gt} based on a pre-trained network (VGG-19 [33]). Let φ_i is the activation map of the i^{th} layer of φ, the perceptual loss is defined as:

$$\mathcal{L}_{perc} = \sum_i ||\varphi_i(I_{edit}) - \varphi_i(I_{gt})|| \quad (9)$$

We exploit the intermediate convolution layer feature maps (*conv_3*, *conv_4* and *conv_5*) of VGG-19 (Pre-trained on ImageNet data [20]) network to get rich structural information and thus helps in recovering plausible structure of the face semantics.

The joint loss function to train the editing module is defined as:

$$\begin{aligned}\mathcal{L}_{comp} &= \lambda_{rc}(\mathcal{L}_{rc} + \mathcal{L}_{perc}) + \lambda_{D^{whole\_region}} \mathcal{L}_{D}^{whole\_region} \\ &\quad + \lambda_{D^{mask\_region}} \mathcal{L}_{D}^{mask\_region} + \lambda_{adv^{whole\_region}} \mathcal{L}_{adv}^{whole\_region} \\ &\quad + \lambda_{adv^{mask\_region}} \mathcal{L}_{adv}^{mask\_region} \quad (10)\end{aligned}$$

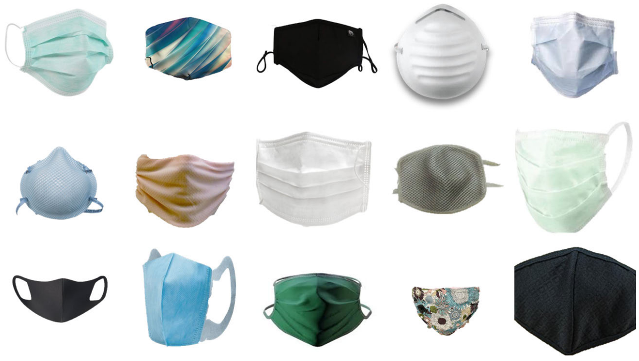


FIGURE 4. Example images of masks used in our synthetic training dataset.

We have set the weight parameters as $\lambda_{rc} = 100$, $\lambda_{D^{whole\_region}} = 0.3$, $\lambda_{D^{mask\_region}} = 0.7$, $\lambda_{adv^{whole\_region}} = 0.3$ and $\lambda_{adv^{mask\_region}} = 0.7$. L_{comp} helps in generating natural looking, structurally consistent and perceptually plausible output.

IV. EXPERIMENTS

In this section, we present synthetic dataset creation, training details of our model and comparison of our method visually and quantitatively with other state-of-the-art image editing approaches. Moreover, in the last part of this section we provide ablation studies of our model.

A. SYNTHETIC DATASET GENERATION

There is no publicly available dataset that contains facial image pairs with and without mask object to train our model in a supervised manner. We construct a synthetic dataset of 10k images using publicly available CelebFaces Attributes Dataset (CelebA) [19]. CelebA is a large-scale face attributes dataset with more than 200K celebrity images. We have used 50 kind of masks of different sizes, shapes, colors and structure in our synthetic dataset. Some of the examples of facial masks in our dataset are shown in Figure 4. To create synthetic samples, we first align the faces using eye-coordinates for all images using dlib [35]. Then we randomly place mask on face using Adobe Photoshop CC 2018. We also generate the corresponding binary map for the mask. Figure 3 shows a couple of examples of our synthetic dataset.

For fair comparison, we have trained current state-of-the-art approaches Iizuka *et al.* [6], Yu *et al.* [9], EdgeConnect [11] and MRGAN [8] using our synthetic dataset. We also provide the object binary map generated by our map module along with input image both at training and inference stages because all these methods assume that object binary map is given.

B. TRAINING DETAILS

For training of the map module, we have fed input image I_{input} into the network and generate a binary map I_{mask_map} that is close to the target binary map I_{tm}. The generated binary map I_{mask_map} along with input image, I_{input}, is then fed into the

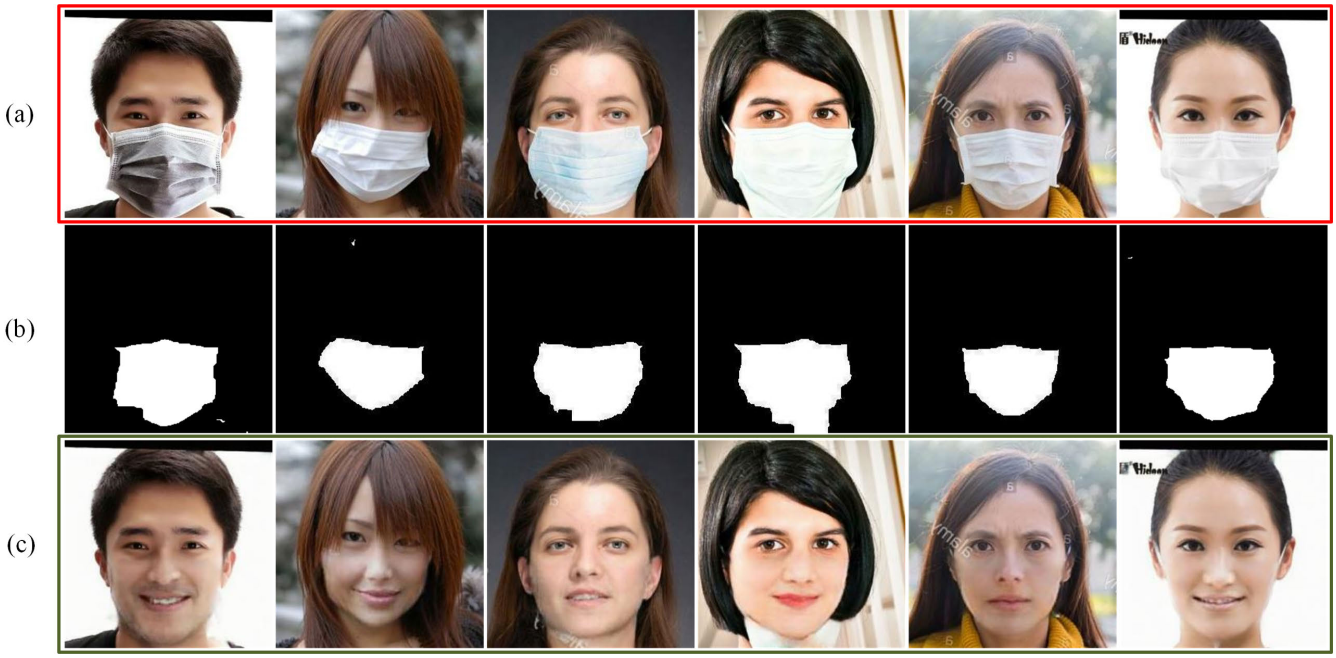


FIGURE 5. Output examples generated by our model for real test samples : a) Input Image, b) Segmentation map, I_{mask_map}, generated by the map module, and c) Output of our model.

editing network and generate the final output I_{edit}. We have implemented our model in tensorflow [36]. Both stages of the model are trained alternatively. We have used 10,000 training samples of size 256 × 256 of our synthetic dataset for training our model with a batch size of 10, and Adam optimizer. We have trained the model for 500,000 iterations.

In the second stage, instead of training the whole network at the same time as done in GLCM [6] and GCA [9], we first train G_{edit} along D_{whole_region} for almost half of the training iterations to generate a reasonable global structure of the face. This helps in getting the actual boundaries of face region. However, it suffers in generating the deep missing region. Once the reasonable global structure of the face is formed, we add D_{mask_region}, to focus more on the missing part for the rest of the training iterations to generate the deep missing semantics more plausibly.

The training details of our model is as follows. To overcome the problem of the editing generator G_{edit} being too weak at the start as compared to the discriminator, we first train the G_{edit} only (no discriminator) for 50,000 iterations and then G_{edit} along with D_{whole_region} for another 200,000 iterations. For the rest of the training iterations, we train the whole network jointly by giving more weight to the D_{mask_region}. This scheme of training helps in providing fair feedback to the editing generator from both discriminators. The whole training procedure takes around about 100 hours using NVIDIA GeForce 2080Ti GPU.

C. COMPARISON AND DISCUSSION

In this section, we analyze results generated by our model and compare with other state-of-the art image editing methods such as Iizuka *et al.* [6], Yu*et al.* [9], EdgeConnect [11] and MRGAN [8] both quantitatively and qualitatively on real world test images.

1) QUALITATIVE COMPARISON

Figure 5 shows the sample generated by our model for real test images. Our test samples contains a lot of diversity in terms of background (blank and wild backgrounds), size, shape, color and structure of masks. In each test image, a mask covers almost half of key facial semantics. As can be seen in Figure 5, our model successfully removes the mask object and generates natural looking outputs with structural consistency. Figure 6 compares our model with the other sate-of-the-art approaches. The results show that our approach successfully removes the mask object and completes the face that looks not only structurally consistent but also naturalistic. Iizuka *et al.* [6], Yu *et al.* [9], EdgeConnect [11] and MRGAN [8] unable to correctly achieve the task. Iizuka *et al.* produce plausible new content in scene images but fail to produce plausible results for face images having missing regions with large structural and appearance variations. The essence of the GCA technique is contextual attention layer which learns to generate missing patches by copying feature information from known background patches. This strategy works well for images where there is high probability of finding same patterns in the neighbouring patches but fails to handle large missing region in facial images (using nose to fill in holes at mouth locations) as shown in third column of Figure 6. Edgeconnect generates better results than GLCM and GCA. However, Edgeconnect's final output depends on the edge map generated by the first stage. For this problem

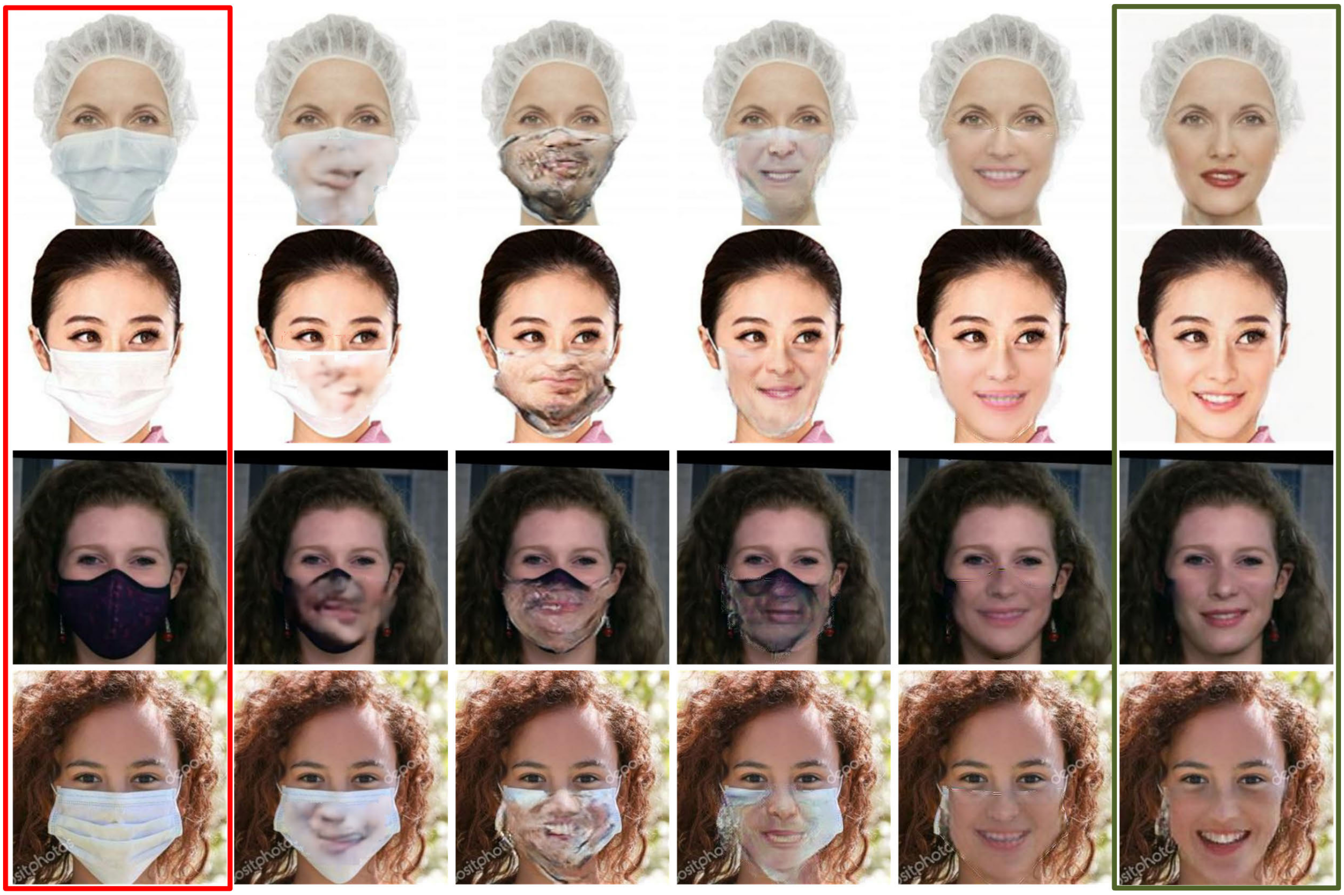


FIGURE 6. Qualitative results comparisons of our model with the other state-of-the art image editing models on real world test images. From left to right: Input image, GLCM [6], GCA [9], EdgeConnect [11], MRGAN [8] and ours result. Note: There is no ground truth since all samples are real world images collected from the Internet.

where the size of the missing region is large, edge generator cannot generate reasonable edges and hence the final outcome of the Edgeconnect network suffers. MRGAN generates only the missing region and keeps the rest of the image as it is. That is why it generates the missing semantics well but also produces some unnatural semantics for the missing region that lies outside the actual boundaries of the face. In summary, our proposed model overcomes the limitations of the other state-of-the-art methods by producing realistic and consistent results regardless of the size, shape, structure and color of facial mask.

2) QUANTITATIVE COMPARISON

We compare the results between our method and the other methods using the following quantitative metrics: 1) Structural SIMilarity (SSIM) [31]; 2) Peak Signal to Noise Ratio (PSNR); 3) Frechet Inception Distance (FID) [37]; 4) Naturalness Image Quality Evaluator (NIQE) [38]; and 5) Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [39]. Table 2 shows quantitative comparison with Iizuka *et al.* [6], Yu *et al.* [9], EdgeConnect [11] and MRGAN [8]. The results show that our model also achieves the best quantitative values. In addation, as can be seen in Table 2, the SSIM value for EdgeConnect and ours model is the same. The reason is that the objective of image editing techniques is to synthesize realistic-looking content rather than the exact same content as the original image. Therefore, we argue that as reported by many other works [7], [40], quantitative analysis may not be the most effective measure of the image editing task.

TABLE 2. Performance comparison of different methods in term of Structural SIMilarity (SSIM), Peak Signal to Noise Ratio (PSNR), Frechet Inception Distance (FID), Naturalness Image Quality Evaluator (NIQE), and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE). The best result of each row are boldfaced.

Methods	SSIM	PSNR	FID	NIQE	BRISQUE
GLCM [6]	0.827	22.40dB	3.651	4.84	38.66
GCA [9]	0.829	18.65dB	4.012	4.68	38.77
EdgeConnect [11]	**0.864**	20.87dB	3.555	4.56	39.60
MRGAN [8]	0.837	25.91dB	3.012	4.86	39.97
Ours	**0.864**	**26.19dB**	**3.548**	**4.46**	**37.85**

3) USER STUDY

We conduct a pilot user study to evaluate our results using perceptual assessment of people. We have asked the total of 20 questions and at every test sample, participants are

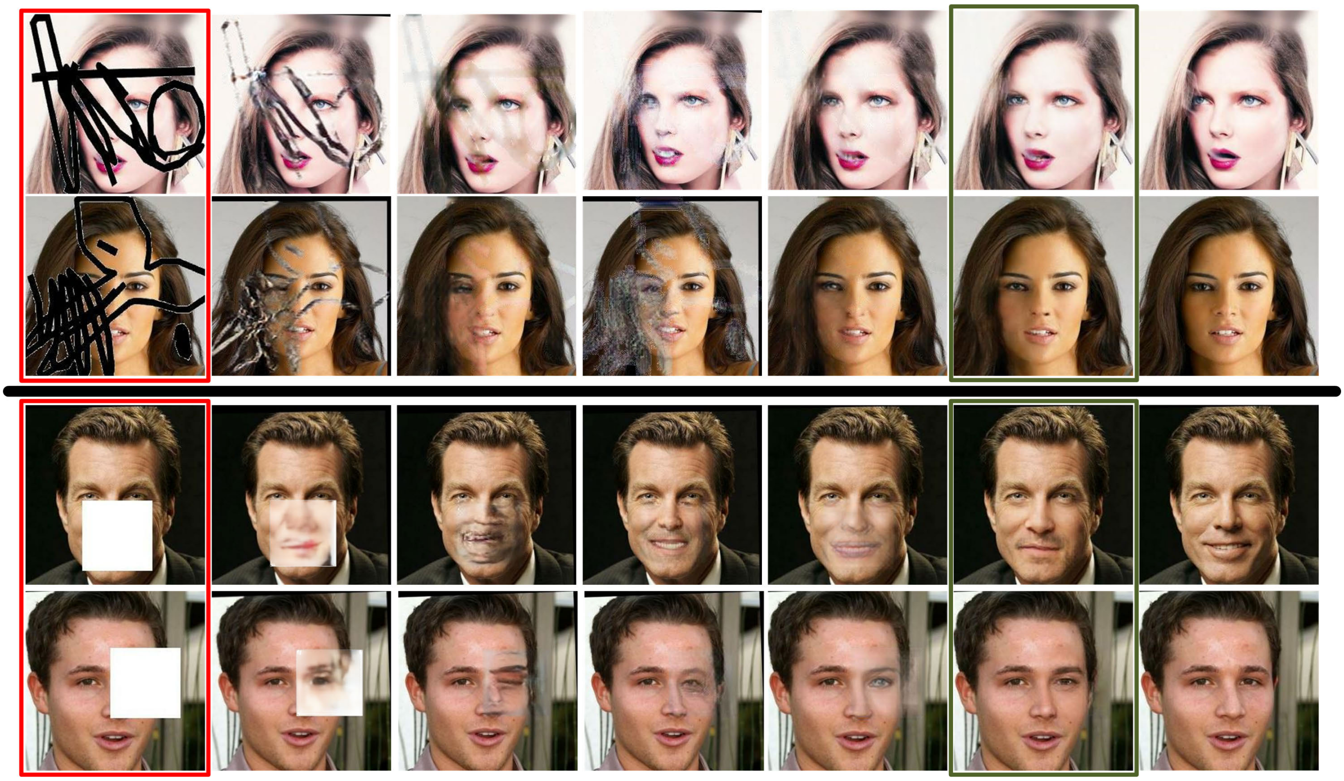


FIGURE 7. **Qualitative results comparisons of our model with other state-of-the art image editing models for irregular and random rectangular missing region. From left to right: Input image, GLCM [6], GCA [9], EdgeConnect [11], MRGAN [8], our result and ground truth. Note: Rectangular missing holes are of size 100 × 100 in 256 × 256 resolution image.**

shown input image along with five randomized options (results produced by Iizuka *et al.* [6], Yu *et al.* [9], [11], MRGAN [8], and our model). We have asked 100 participants to choose one option out of the five that have effectively removed the mask and complete the image while retaining natural look and structure of the face. Our results got 76 votes, MRAGN [8] and EdgeConnect [11] earned 10 and 7 votes, respectively, while Iizuka *et al.* [6], Yu *et al.* [9] were voted by 5 and 2 voters.

4) ADDITIONAL RESULTS

We have retrained our model and the other state-of-the-art representative models for inpainting irregular and random rectangular missing holes. In case of the other state-of-the-art models, we provide the object binary map generated by our map module both at training and inference stages. As can be seen in Figure 7, our model completes the face images while retaining the naturalness and structure of the faces comparably well to the other state-of-the-art approaches. Moreover, in Table 3, we can see the quantitative comparison for both irregular and rectangular missing region. The results show that our model achieves better quantitative performance in most of the cases for completing diverse missing regions in facial images.

5) LIMITATIONS

Although our model can handle removal of mask objects of various shapes, size, color and structure, there are some examples as can be seen in Figure 8 where our model fails to completely remove the mask object. Common failure cases occur when the map module is unable to produce a reasonable segmentation map of the mask object. This happens when mask objects are very different than those in our synthetic dataset in terms of both shape and structure. As can be seen in the first couple of rows of Figure 8, the shape, color and structure of the mask objects are totally different than the mask types we used in our synthetic dataset, failing to detect them properly. In the third row, the network is unable to detect the whole mask region due to complex mixture of colors. The network failed to detect the part where its color is similar to face texture because it was considered part of the face.

D. ABLATION STUDY

1) ROLE OF USING TWO DISCRIMINATORS

We investigate the effectiveness of using one discriminator at a time and using both (D_{whole_region} and D_{mask_region}) by gradually adding them to the model. The first column of Figure 9 shows the result of using only D_{mask_region} with the rest of the setting same as our model. As this setting will only focus on generating the affected region and keep the rest of the image same as original input image, it produces good semantics at the missing area, e. g., mouth and down part of the nose. However, as can be seen in Figure 9 (b), it mixes the chin with the neck (specially neck area covered by the object) by considering it as part of the chin. This setting generates

TABLE 3. Performance comparison of different methods for completing irregular and rectangular missing regions in term of Structural SIMilarity (SSIM), Peak Signal to Noise Ratio (PSNR), Frechet Inception Distance (FID), Naturalness Image Quality Evaluator (NIQE), and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE). The best result of each column row are boldfaced.

Methods	Irregular missing region					Rectangular missing region				
	SSIM	PSNR	FID	NIQE	BRISQUE	SSIM	PSNR	FID	NIQE	BRISQUE
GLCM [6]	0.819	17.439dB	11.844	4.929	36.997	0.894	18.787dB	39.088	4.735	39.521
GCA [9]	0.702	18.610dB	6.165	7.727	32.410	0.924	26.913dB	24.006	**3.828**	34.026
EdgeConnect [11]	**0.837**	25.593dB	4.919	5.961	27.243	0.926	27.373db	8.910	4.087	29.661
MRGAN [8]	0.821	28.097dB	4.967	4.796	29.300	0.825	28.214dB	6.740	3.944	29.115
Ours	0.824	**28.099dB**	**4.965**	**3.855**	**27.008**	**0.928**	**28.241dB**	**6.102**	3.898	**29.087**

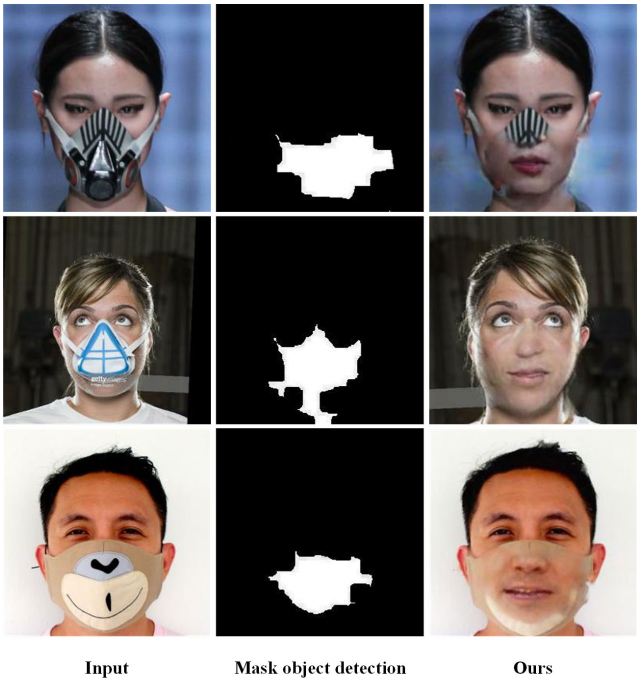


FIGURE 8. Failure cases of our model. The map module affects the editing module by not properly detecting the mask object.

the worst results when mask color is similar to the neck color. To overcome the problem, instead of using D_{mask_region}, we have additionally used D_{whole_region} with the rest of the setting same as our original model. We drop the D_{mask_region} and use only D_{whole_region} along with the editing generator. The second column of Figure 9 shows results produced by this setting. We can see that it produces more consistent structure of the face and does not mix chin with neck but this setting is incapable of synthesizing plausible content in the deep region of the missing part: it produces teeth that look neither symmetric nor natural. In order to generate plausible content in the missing region that is consistent with the rest of the face, we have used both discriminators along the editing generator by adding each discriminator gradually to the network as stated earlier in training part of the experiment section. The last column of the Figure 9 shows that our training strategy of two discriminators not only generates plausible contents under the large missing region but also recovers correct semantic structure.

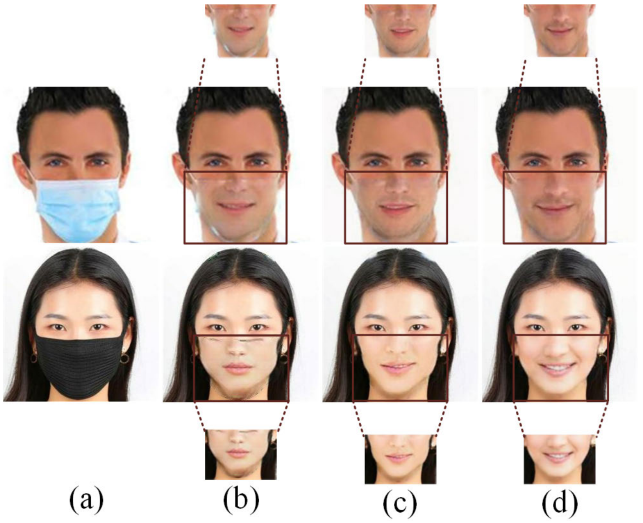


FIGURE 9. Effect of using multiple discriminator (D_{whole_region} and D_{mask_region}). a) Input image with Mask. b) Result of our model using only D_{mask_region}. c) Result of our model using only D_{whole_region}. d) Result of our model using both discriminator and the training scheme as explained in training details section.

2) EFFECT OF MAP MODULE

We have dropped the map module part and only used the editing network to validate the effect of I_{mask_map} on object removal and image editing. For this, we have only fed I_{input} (image with mask) into the editing network generator while the rest of the model is kept the same as our baseline network. Figure 10 shows that without using mask segmentation in the image editing network produces irregular structure of the face. For example, in first row of Figure 10 (b), the texture of the generated region is different from the rest of the face, while in second and third rows, we can see that lips are mixed with each other and boundaries of the chin looks very unnatural. On the other hand, we can see in Figure 10 (c) that using the segmentation map helps not only recover the correct texture of the damaged region but also generate sharp boundaries of the chin part covered by mask. This shows that mask segmentation provides enough information about where the object pixels are and makes the task easy for the image editing network. Hence, using mask segmentation along the input image for editing network results in more accurate object removal and realistic face image editing.

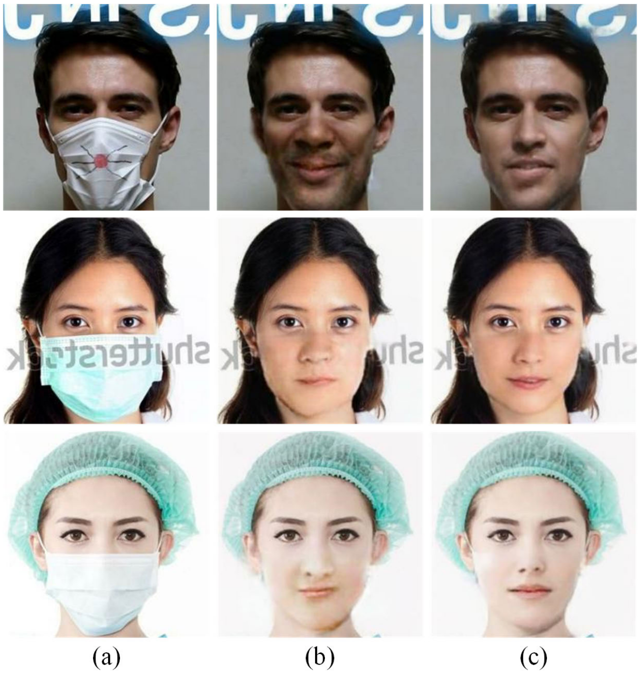


FIGURE 10. Effect of providing mask segmentation along with input image to the completion network. a) Input image. b) Results of our model without using mask segmentation at the editing module input c) Results of our model using both mask segmentation and input image at the editing module input.

V. CONCLUSION

In this work, we have proposed a novel method for interaction-free large object removal from facial images, focusing on mask object. For image completion, we have employed GAN based image inpainting through image-to-image translation approach to produce plausible results. We have shown that the proposed training scheme of two discriminators for gradually learning global coherency and deep missing region is quite effective in producing realistic and structurally consistent outputs. Both qualitative and quantitative comparison show that our model is capable of producing high perceptual quality results for large missing hole in facial images as compared to other state-of-the art image editing methods.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[2] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[3] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen, "Image melding: Combining inconsistent images using patch-based synthesis," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–10, Jul. 2012.

[4] J. Hays and A. A. Efros, "Scene completion using millions of photographs," *ACM Trans. Graph.*, vol. 26, no. 3, p. 4, 2007.

[5] J.-S. Park, Y. Hwa Oh, S. Chul Ahn, and S.-W. Lee, "Glasses removal from facial image using recursive error compensation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 805–811, May 2005.

[6] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Jul. 2017.

[7] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6721–6729.

[8] M. K. J. Khan, N. Ud Din, S. Bae, and J. Yi, "Interactive removal of microphone object in facial images," *Electronics*, vol. 8, no. 10, p. 1115, Oct. 2019.

[9] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[10] K. Javed, N. Ud Din, S. Bae, and J. Yi, "Image unmosaicing without location information using stacked GAN," *IET Comput. Vis.*, vol. 13, no. 6, pp. 588–594, Sep. 2019.

[11] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "EdgeConnect: Generative image inpainting with adversarial edge learning," 2019, *arXiv:1901.00212*. [Online]. Available: http://arxiv.org/abs/1901.00212

[12] Y. Song, C. Yang, Y. Shen, P. Wang, Q. Huang, and C.-C. J. Kuo, "SPG-NET: Segmentation prediction and guidance network for image inpainting," in *Proc. BMVC*, 2018, pp. 1–14.

[13] J. Wang, K. Lu, D. Pan, N. He, and B.-K. Bao, "Robust object removal with an exemplar-based image inpainting approach," *Neurocomputing*, vol. 123, pp. 150–155, Jan. 2014.

[14] B. Hwang, J. Park, Y. Ham, S. Bang, J. Lee, M. Roh, J. Moon, and S. Lee, "Design and construction of a korean face database for research and development," in *Proc. Korea Inf. Sci. Soc. Workshop Comput. Vis. Pattern Recognit.*, Seoul, South Korea, 2000, pp. 161–163.

[15] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and Web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, 2008.

[16] R. R. Shetty, M. Fritz, and B. Schiele, "Adversarial scene editing: Automatic object removal from weak supervision," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7706–7716.

[17] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. Lawrence Zitnick, "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*. [Online]. Available: http://arxiv.org/abs/1504.00325

[18] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[19] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Apr. 2015.

[21] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, "What makes Paris look like Paris?" *Commun. ACM*, vol. 58, no. 12, pp. 103–110, Nov. 2015.

[22] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3911–3919.

[23] G. Dong, W. Huang, W. A. P. Smith, and P. Ren, "A shadow constrained conditional generative adversarial net for SRTM data restoration," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111602.

[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[25] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Munich, Germany: Springer, 2015, pp. 234–241.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE CVPR*, Jun. 2018, pp. 7132–7141.

[30] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[32] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.

[34] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 694–711.

[35] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," CMU School Comput. Sci., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-16-118, 2016.

[36] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, vol. 16. 2016, pp. 265–283.

[37] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.

[38] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[39] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/Referenceless image spatial quality evaluator," in *Proc. Conf. Rec. 45th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2011, pp. 723–727.

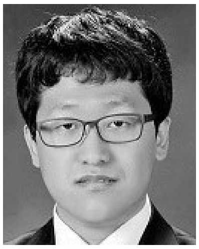
[40] K. Javed, N. U. Din, S. Bae, R. S. Maharjan, D. Seo, and J. Yi, "UMGAN: Generative adversarial network for image unmosaicing using perceptual loss," in *Proc. 16th Int. Conf. Mach. Vis. Appl. (MVA)*, May 2019, pp. 1–5.

NIZAM UD DIN received the B.Sc. degree in electrical (computer) engineering from the COMSATS Institute of Information Technology, Pakistan, in 2013, and the M.Sc. degree in computer engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2016. He is currently pursuing the Ph.D. degree with the Computer Vision Laboratory, Department of Electrical and Computer Engineering, Sungkyunkwan University, South Korea.

He joined Quaid-i-Azam University, Islamabad, Pakistan, as a Visiting Faculty member. He is mainly interested in deep learning, especially applied to computer vision. His awards and honors include University Scholarship for B.Sc. Electrical (Computer) Engineering from the COMSATS Institute of Information Technology, Pakistan, in 2009, and Higher Education Commission HRDI- FACULTY DEVELOPMENT OF UESTPS-UETS Scholarship for Ph.D., from HEC, Pakistan, in 2017.

KAMRAN JAVED received the B.Sc. degree (Hons.) in electronic engineering and the M.Sc. degree in computer engineering from the University of Engineering and Technology (UET), Taxila, Pakistan, in 2012 and 2014, respectively, and the Ph.D. degree in electronic and computer engineering from Sungkyunkwan University, South Korea, in 2020.

He was a Lecturer with the Electronic Engineering Department, University of Engineering and Technology, from 2013 to 2016. His research interests are focused on generative adversarial networks and its application to computer vision for image unmosaicing and object removal. His awards and honors include the Award of Honors in B.Sc. degree from UET, in 2012, the University Scholarship for his M.Sc. degree from UET, in 2012, and the Higher Education Commission Scholarship for his Ph.D. degree from HEC, Pakistan, in 2016.



SEHO BAE received the B.S. degree in electronic and electrical engineering from Sungkyunkwan University, Suwon, South Korea, in 2015, where he is currently pursuing the integrated Ph.D. degree in electrical and computer engineering. He has been a member of the Computer Vision Laboratory, since 2015. His current research interests include cross-modal image matching and cross-modal image synthesis using generative adversarial networks.

JUNEHO YI received the B.S. degree from Seoul National University, South Korea, in 1985, the M.S. degree from Pennsylvania State University, University Park, PA, USA, in 1987, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 1994, all in electrical engineering. In 1989, he was a Research Scientist with the Samsung Advanced Institute of Technology. From 1994 to 1995, he was a Research Scientist with the University of California at Riverside. From 1995 to 1996, he was a Senior Research Scientist with the Korea Institute of Science and Technology, Seoul, South Korea. Since 1997, he has been with Sungkyunkwan University, South Korea, where he is currently a Professor with the School of Electronic and Electrical Engineering. His pioneering works include masked fake face detection and depth filtering using parameterized structured light imaging. His research interests are broadly in the areas of computer vision and statistical pattern recognition.