International Journal of Communication Networks and Information Security
2024, 16(4)
ISSN: 2073-607X, 2076-0930
https://https://ijcnis.org/ **Research Article**

Deep Transfer Learning for Masked Face Reconstruction and Hybrid DCNN-ELM Framework for Recognition

Chandni Agarwal[1], Anurag Mishra[2], Charul Bhatnagar[3]

[1]GLA University, Mathura, Uttar Pradesh – 281406, India
ORCID ID: 0000-0001-8387-5969
[2]DeenDayal Upadhayay College, Delhi University, New Delhi-110078, India
[3]GLA University, Mathura, Uttar Pradesh – 281406, India

Email: [1]Chandni1972@gmail.com, [2]Anuragm1967@ddu.du.ac.in, [3]charul@gla.ac.in

ARTICLE INFO

Received: 28 Apr 2024
Accepted: 05 Sep 2024

ABSTRACT

Facial reconstruction has always been a pivotal aspect of medical and forensic science. The increasing use of face masks in recent years has posed new challenges, making traditional facial recognition techniques less effective. To address this, our research explored innovative methods for reconstructing faces from images obscured by masks. We focused on post mask face reconstruction and facial recognition using cutting-edge techniques. We assess the effectiveness of three key unmasking algorithms: edgeconnect (EC), gated convolution (GC), and hierarchical variational vector quantized autoencoders (HVQVAE). Using two synthetic face datasets, MaskedFace-CelebA and MaskedFace-CelebAHQ, we rigorously evaluate the quality of the reconstructed faces based on metrics such as the PSNR, SSIM, UIQI, and NCORR. Among these, the Gated Convolution algorithm stands out as the superior choice in terms of image quality. For facial recognition, we employ a novel hybrid framework that combines a deep convolutional neural network and an extreme learning machine (DCNN-ELM). We tested five classifiers (Vgg16, Vgg19, ResNet50, ResNet101, and ResNet152) in combination with ELM and a support vector machine (SVM). Our comprehensive ablation study revealed that ResNet152 combined with ELM achieved the best performance, with a facial recognition accuracy of 60.9%, suggesting that the reconstructed faces were of high quality. Our paper presents a novel approach to image classification utilizing five classifiers within the DCNN+ELM hybrid framework and provides a complete ablation study of these classifiers. This research underscores the importance of face reconstruction in the current field and its potential to enhance facial recognition techniques.

Keywords: Image Inpainting, GAN, Autoencoders, Deep Learning, Face Reconstruction, Face Recognition

INTRODUCTION

The advent of face masks as common accessories in daily life has introduced a unique challenge to the field of facial recognition. Traditional systems, which rely on facial landmarks for identification, are often stumped when these key features are obscured by masks[1]. This issue is particularly acute in forensics, where accurate identification is paramount. In response, our paper proposes a comparative study in which we reconstruct the faces of masked individuals for recognition using three deep transfer learning methods. This research holds significant practical value, especially in forensics, and addresses several key issues: (1) determining if the mask is correctly worn, (2) accurately identifying masked versus unmasked faces, and (3) matching masked faces with a database of masked faces for face recognition. While existing facial recognition models perform well for nonmasked faces, they struggle with occluded or covered faces. Therefore, reconstruction is crucial for accurate facial recognition outcomes. Notably, masked face reconstruction for face recognition is in its initial phase, and limited research has been conducted on this topic. Current studies focus on determining whether a mask is correctly worn, identifying masked or unmasked faces, or matching masked faces with their ground truth images within databases. The last objective is particularly relevant for image forensics, which has broader applications. Consequently, the reconstruction or regeneration of faces is a crucial area of research in this domain. In the proposed work, we demonstrate the reconstruction of an unmasked face from a masked face using different generative modeling techniques. The experiments are carried out using image inpainting or an image completion mechanism for reconstructing the face (Figure 1) via deep learning methods such as the generative adversarial network (GAN) [3,5,14,15,33] and the variational autoencoder (VAE) [4,7,8].

Image inpainting[10] is a well-established image processing technique that fills missing or corrupted areas with diverse content; finds applications in recolouring, restoration, and distortion removal; and extends to tasks such as rotation, stitching, retargeting, compression, and regeneration. While efficient automatic image completion algorithms exist, reconstructing large masked areas remains challenging. Due to advancements in computer vision, deep learning-based approaches [7,27] excel in capturing intricate high-level semantics, yielding markedly improved results. In this work, state-of-the-art algorithms for face regeneration and output comparison are employed.

Fig. 1 First Row shows Ground Truth, Second Row shows masked faces, Third Row - Green box shows the generated images using Proposed HVQVAE based transfer learning model.

To accomplish this, we used our own synthetic masked face datasets, namely, MaskedFace-CelebA [30] and MaskedFace-CelebA-HQ [30], created using the MaskTheFace [12] tool over the benchmark face datasets CelebA [5] and CelebA-HQ [14].

First, the faces are reconstructed using the edge-connect model, which is based on generative image inpainting with adversarial edge learning and was recently proposed by Nazeri K. et al. in 2019 [11]. This model has limited applications. It hallucinates edges in the missing regions through a two-stage process. The first stage focuses on generating edges in the missing regions, while the second stage uses these generated edges in an image completion network to estimate the RGB intensities of the missing regions. Another model tested for face reconstruction is free-form image inpainting with the gated convolution model, proposed by Yu et al. [15]. This model aims to reconstruct faces after removing occlusions, which may be rectangular in shape. It employs a dual-phase image inpainting network. The initial phase uses a dilated convolutional network trained with reconstruction loss to approximate the missing regions. The subsequent stage incorporates contextual attention to ensure the spatial coherence of attention. These deep learning models utilizing GANs have been proposed for various applications, such as removing occluded areas in faces and completing images of buildings. In this work, we use these models to generate nonmasked faces from masked faces.

Recent studies focusing on deep learning approaches for image inpainting commonly utilize encoder–decoder architectures trained with a blend of reconstruction and adversarial losses [24, 4, 9]. The latest model assessed for unmasking faces employs the hierarchical VQ-VAE framework proposed by Peng J. et al. in 2021 [16]. This innovative model employs a dual-stage process for diverse image inpainting using autoencoders. Initially, multiple coarse results are generated, each with distinct structures. Subsequently, each coarse result undergoes refinement independently through texture augmentation. The hierarchical architecture of this model segregates structural and textural data, facilitated by vector quantized autoencoders, enabling autoregressive modeling of discrete distributions over structural information. Originally designed for multiple-solution inpainting, the HVQVAE method uses an autoregressive distribution to generate varied structures, followed by synthesizing image textures to maintain consistency with the generated structure. While this model is evaluated for reconstructing masked faces, previous applications include filling in missing regions (blocks) in images of buildings, individuals, and other contexts.

This paper is structured as follows: Section 2 covers previous work in the domain of face reconstruction, focusing on GAN-based models and image inpainting techniques. Section 3 outlines the motivation and contributions of our study, while Section 4 details the experimental setup. Section 5 delves into the experimental results and analysis, and Section 6 introduces our novel approach, utilizing five classifiers to showcase the comparative performance of facial recognition in terms of accuracy when distinguishing between ground truth and reconstructed faces. Finally, Section 7 provides insight into the ablation study, and Section 8 provides concluding remarks for our study on face reconstruction and recognition employing a deep transfer learning approach.

RELATED WORKS

2.1 Face reconstruction

Different deep learning-based painting techniques are used to achieve aesthetically realistic results, and various recent studies are discussed in this section. The earlier context encoder [2] was built on deep learning-based GANs, where the resulting output image of an input image with missing parts was composed of oversmoothed or fuzzy patches due to an information bottleneck in the channelwise fully connected layer. In 2017, Yang, C. et al. [23] introduced an updated version of the context encoder by utilizing neural style transfer learning. The

proposed technique helps improve the textural details of the generated pixels, but at the same time, a longer training period is needed to reach real-time performance.
In 2017, Iizuka et al. [24] suggested both global and local discriminators as adversarial losses to address this problem by replacing the channelwise fully linked layer with a sequence of dilated convolution layers. This approach resulted in extremely sparse filters, allowed us to accommodate a wide range of input image sizes and used multiscale discriminators to enhance the textural details of the final images. However, compared to natural images, filling in missing parts of images did not produce satisfactory results. In 2018, Liu et al. [5] proposed a method for normalizing convolution weights by the window mask area, and he was the first to suggest a model for dealing with irregular gaps in inpainting.

Adding to the research landscape, Yu et al. [15] introduced a novel two-stage image inpainting network. The initial stage employed a dilated convolutional network trained with reconstruction loss to provide an initial estimation of the missing regions. Subsequently, the integration of contextual attention in the second stage aimed to promote spatial coherence of attention. Addressing mask removal, Boutros et al. [37] introduced an embedding unmasking model that utilized a feature embedding extracted from the masked face, generating a new feature embedding to that of an unmasked face. Moreover, Din et al. [35,36] employed GAN-based image inpainting through an image-to-image translation approach to facilitate the automatic removal of face masks.
In addition to the aforementioned studies, we previously conducted a comprehensive comparative analysis of learning and nonlearning image inpainting models, as described in our prior study [30]. In our current research, we compared the edgeconnect method [11] and the gated convolution method [15] for the purpose of removing COVID-type face masks and reconstructing unmasked faces. Our experimental study, along with quantitative analysis using full-reference image quality assessment metrics, demonstrates the superiority of the Gated Convolution method over the Edge Connect method.

Recent research has indicated that autoencoder-based models may yield suboptimal results in regard to face or image reconstruction in cases involving occlusion removal. Therefore, we also evaluated the performance of the HVQVAE model [16] in reconstructing unmasked faces using the same dataset. This choice is motivated by our desire to explore alternative approaches in light of the challenges faced in previous research, as highlighted in our earlier work[30]. Zheng et al. [27] suggested a VAE-based model with two parallel routes to obtain numerous inpainting solutions, which trades off between recreating ground truth and retaining the diversity of inpainting results. Zhao et al. [7] presented a VAE-based model that improves diversity by using instance photos. However, because these approaches fail to efficiently segregate structural and textural information, deformed structures and/or hazy textures frequently arise. To solve the issue of deformed structures and/or hazy textures, J. Peng et al. [16] suggested an HVQVAE-based model that leverages the hierarchical VQVAE architecture to separate structural and textural information. In addition, VQVAE's vector quantization allows for autoregressive modeling of the discrete distribution over structural information, as well as two feature losses, to increase structure coherence and texture realism.

2.2 VQ-VAE and Autoregressive Networks

The vector quantized variational autoencoder (VQ-VAE) [17] is a discrete latent VAE model that models discrete latent variables using vector quantization layers. Because latent variables are discrete, a sophisticated autoregressive network such as PixelCNN [18,19,20] can model them without worrying about the posterior collapse problem [17]. Razavi et al. [21] proposed a hierarchical VQ-VAE that uses a hierarchy of discrete latent variables to separate structural

and textural information, followed by two PixelCNNs to model structural and textural information. However, the PixelCNNs for image production are conditioned on the class label, whereas there is no class label in the image inpainting task. Furthermore, due to the lossy nature of VQ-VAE, the generated textures of the PixelCNNs lack fine-grained features. This approach frees the generation model from simulating insignificant data and prevents the inpainting model from generating realistic textures consistent with the known regions. As a result, PixelCNNs are unsuitable for picture inpainting.

2.3 Hierarchical VQVAE

Inspired by the hierarchical VQ-VAE paradigm [21], Peng, J[16] suggested a model for producing different images using HVQVAE in which the hierarchical encoder translates ground truth into structural and textural elements. Two vector quantization layers convert these features to discrete features. The codebook of each vector quantization layer has K = 512 prototype vectors, with a vector dimensionality of D = 64. As a result, each feature vector is replaced by the prototype vector that is the closest in Euclidean space. Finally, using these two sets of discrete characteristics, the decoder reconstructs an image. The straight-through gradient estimator [22] is used to backpropagate the gradient of the reconstruction loss through vector quantization. The exponential moving average of the encoder output is used to update the prototype vectors in the codebook. The size of the structural features in the 256 × 256 images is 3232, and the size of the textural characteristics is 6464. Textural features model local information such as details and textures, while structural features model global information such as shapes and colors. This model was tested in this study for masked face reconstruction.

1. Motivation and contribution of the proposed work

It is clear from the discussion in Sections 1 and 2 that face reconstruction from masked faces using existing deep learning models is a difficult task. However, in this paper, we report that these methods have achieved a certain degree of success in reconstructing these faces from those faces created after masking the same ones as CelebA-HQ. For this purpose, we have evolved a strategy that has resulted in the following major contributions:

1. **Assessment of Existing Models:** We evaluated the performance of the edge-connect transfer learning model and free-form image inpainting with gated convolution on a novel dataset, consisting of 9,622 images from MaskedFace-CelebA and 9,758 images from MaskedFace-CelebA-HQ. The findings, detailed in Sections 6 and 7, provide new insights into these models' effectiveness for masked face reconstruction.
2. **Enhancements to the HVQVAE Framework:** We tested the HVQVAE framework for masked face reconstruction using transfer learning. Initial results were suboptimal, prompting us to introduce targeted modifications that significantly improved reconstruction quality. These enhancements are thoroughly analyzed in Section 6.
3. **Comparative Analysis of Transfer Learning Methods:** Through extensive experiments, we found that the Gated Convolution method outperformed other models tested on the same datasets. This result represents a significant step forward in developing robust techniques for masked face reconstruction.
4. **Advanced Model for Face Recognition:** We employed five DCNN-TL models—Vgg16, Vgg19, ResNet50, ResNet101, and ResNet152—for recognizing reconstructed faces. The ResNet152 model, combined with an ELM classifier using sigmoid activation, demonstrated the best performance, as confirmed by a comprehensive ablation study.

These contributions advance the application of deep learning techniques in masked face reconstruction, offering new approaches and improvements that are valuable for future research in image forensics.

4. Experimental Setup

4.1 Datasets

To carry out the experiments effectively, we used our own two masked face datasets, namely, the MaskedFace-CelebA (Fig. 2) and MaskedFace-CelebA-HQ (Fig. 3) datasets, created using the CelebA[5] and CelebA-HQ[14] benchmark face datasets, respectively, using the MaskTheFace[12] tool. These masked face datasets are aligned with nonmasked faces to make the datasets aligned.

Fig. 2 Synthetic mask application (original face and masked face) for the CelebA dataset

Fig. 3 Synthetic Mask Application (Original Face & Masked Face) for the CelebA-HQ Dataset

To create synthetic samples, original images of the CelebA-HQ dataset of size 1024 × 1024 were scaled down to 256 × 256, and for the CelebA dataset, images were scaled up to 256 ×256 from size 178 × 218 for further processing to fit the algorithm's requirements (Fig. 4).

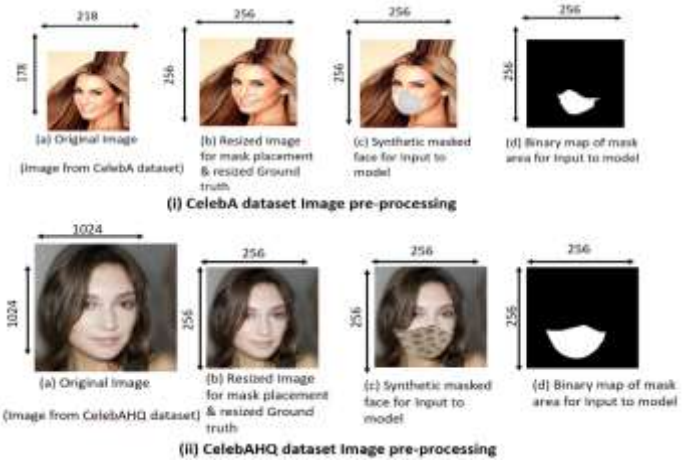

Fig. 4 Preprocessing of the image dataset (data resizing, mask placement & binary map segmentation)

The first dataset included 37790 images, and the mask was placed over 36910 images successfully in the Masked Face-CelebA dataset. The second synthetic masked face dataset was created with 29571 images using the original CelebA-HQ [14] large-scale face image dataset with 30K high-resolution face images.

4.2 Face Reconstruction Phase

Previous studies in face recognition and computer vision have developed many state-of-the-art models capable of identifying faces with high accuracy. However, these models struggle to maintain performance when faces are covered by masks. This paper explores the use of masked face reconstruction to improve unmasked face recognition using existing deep learning models. We focus on three transfer learning models commonly used for image restoration tasks, but their effectiveness for facial image reconstruction varies. These models are discussed as follows:

4.2.1 Edge Connect: Generative Image Inpainting with Adversarial Edge Learning

The edge-connect technique, introduced by Nazeri et al. [11], uses a "lines first, color next" approach inspired by artistic methods. It employs an edge generator to predict missing edges in images based on existing edges and grayscale pixel data, followed by an image completion network that fills in missing regions with color and texture details. The edge-connect architecture features a dual-stage edge-to-image network with two generators and two discriminators, utilizing adversarial, feature matching, style, perceptual, and L1 reconstruction losses to train the model. The edge generator aims to replicate edge maps identified by the edge discriminator, enhancing consistency in feature representation. This method effectively reconstructs images with missing portions, such as those in masked face datasets where detailed reconstruction is essential. The system takes an 8-bit binary mask and a 24-bit overlayed mask image as input and outputs the reconstructed image. Figure 5 illustrates the EdgeConnect [11] results, showing the reconstructed images generated from the face-with-mask overlay and binary map inputs, leading to the final output displayed in Figure 6.

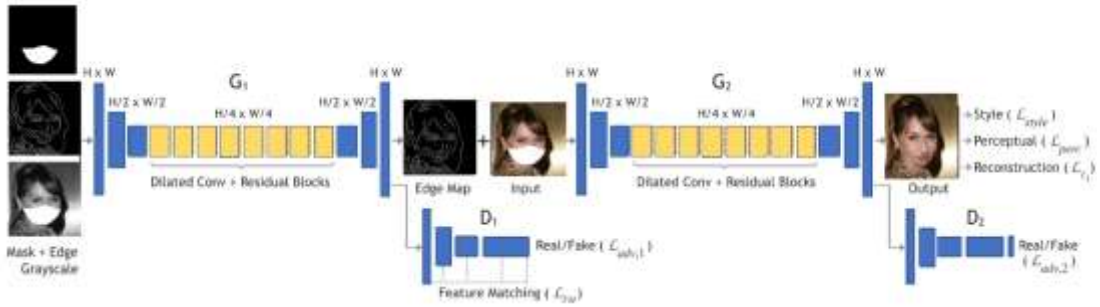

Fig. 5 Network architecture – edgeconnect [11]. The incomplete grayscale image, edge map, and mask are the inputs of G1 to the inputs of G1 to predict the full edge map. The predicted edge map and incomplete color image are passed to G2 to perform the inpainting task.

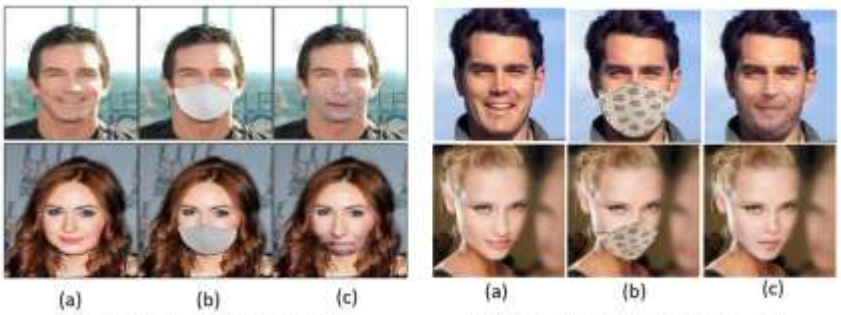

Fig. 6 Face reconstruction using the edgeconnect [11] model evaluated on the MaskedFace-CelebA and MaskedFace-CelebA datasets: (a) ground truth, (b) masked image, and (d) reconstructed face

4.2.2 Free-form Image Inpainting with Gated Convolution

Yu, Jiahui et al. [15] developed a generative image inpainting system for completing images with a free-form mask and user guidance. This system combines elements from previous methods, including the contextual attention (CA) layer from DeepFill v1, user guidance inspired by edge-connect, and gated convolution (GConv), an adaptation of partial convolution. In GConv, the mask update is handled by a learnable gating mechanism for subsequent convolution layers (see Figure 7). Known as DeepFill v2, this model achieves high-quality free-form inpainting, outperforming earlier methods with its two-stage, coarse-to-fine network structure: the first generator performs coarse reconstruction, while the second refines the image.

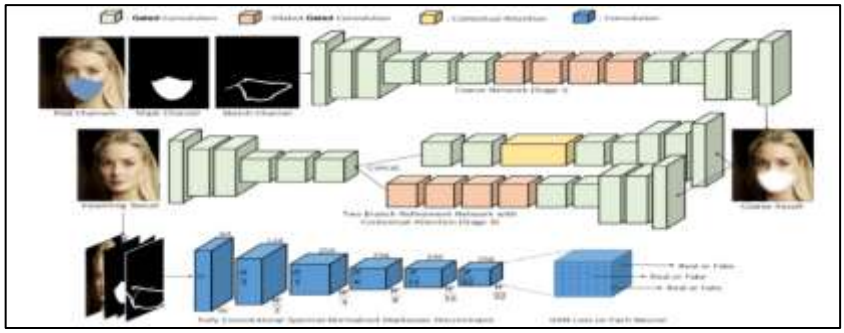

Fig. 7 Network Architecture : Free form Image Inpainting with Gated Convolution (As per Original literature)

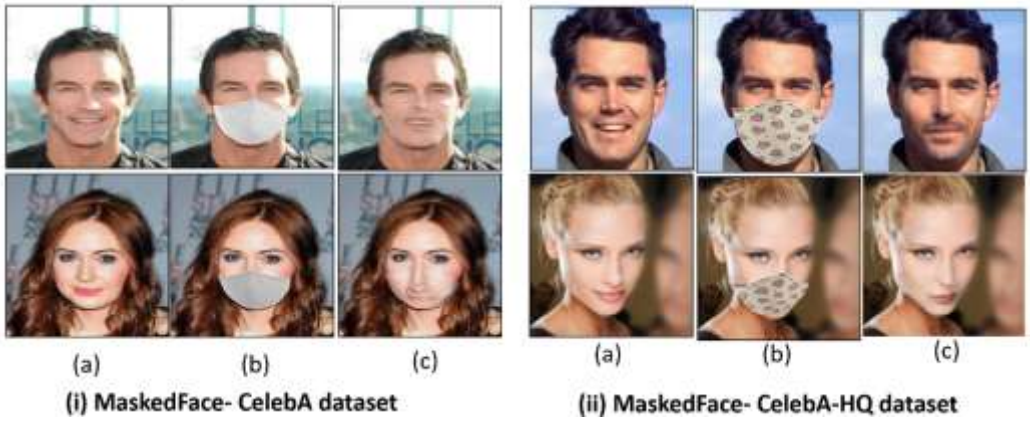

Fig. 8 Face reconstruction using the Gated Convolution [11] model evaluated on the MaskedFace-CelebA and MaskedFace-CelebAHQ datasets: (a) ground truth, (b) masked image, and (c) reconstructed face

The network is trained using L1 and GAN losses. We evaluated this architecture on our masked face dataset after preprocessing, assessing 9,621 images from MaskedFace-CelebA and 9,883 from MaskedFace-CelebA-HQ. The model showed superior performance on the MaskedFace-CelebA-HQ dataset (Figure 8), benefiting from a pretrained model based on CelebA-HQ. Detailed results are discussed in Section 5.

4.2.3 Generating a Diverse Structure for Image Inpainting with a Hierarchical VQ-VAE

This is the third approach, which relates to generating diverse structures for image inpainting with hierarchical VQ-VAE on our masked dataset. This approach was originally proposed by Peng, J.et al. [16]. The authors proposed a multiple-solution image inpainting technique that uses a two-stage model for diverse inpainting,

where the first stage generates multiple coarse results, each of which has a different structure using an autoregressive distribution over latent variables, and the second stage refines each coarse result separately by splitting structural and textural features. To improve the structural coherence and texture realism, two feature losses were also proposed. Figure 9 shows the process flow diagram of the model proposed by Peng J. et al. [16] on the selected images from our dataset. In Figure 9, a three-layer framework is demonstrated where the top layer reconstructs the image from structural and discrete textural features using HVQVAE, the middle layer models the conditional distribution using an autoregressive network and the bottom layer generator synthesizes the image texture given the discrete structural features.

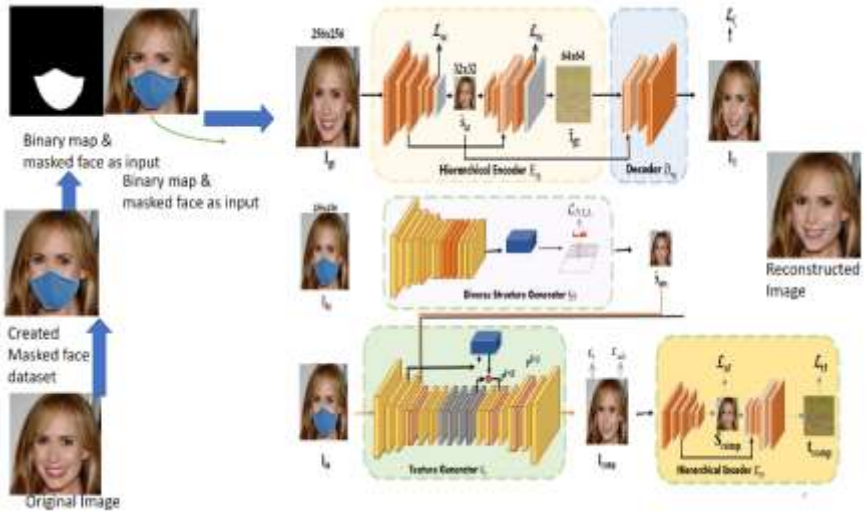

Fig. 8 Network architecture: generating diverse images using HVQVAE[16] (as per the original literature)

On our own dataset, we experimented with this architecture for regenerating faces from mask-covered faces. Figure 10 shows the outcome of the existing HVQVAE algorithm for masked face reconstruction.

As the results shown in Fig. 10 are not promising, we propose a new algorithm to improve facial reconstruction based on the HVQVAE[16] model. The proposed algorithm is based on the fine-tuning of the hyperparameters in terms of changing the weights of the commitment loss in the vector quantizer and adding or modifying the existing code for testing the model (the HVQVAE model) to reshape and normalize the masked area to be reconstructed by adding the masked image and mask of the image. The intensity of the pixels is reduced to 255 wherever it is higher than 255 to reach the maximum intensity. The modified algorithm yields improved results in terms of the PSNR and SSIM, which are image quality assessment metrics. Figure 10 shows the qualitative and quantitative results of the outputs generated using the HVQVAE and modified HVQVAE models. The modified HVQVAE model yields better outputs for the same faces reconstructed by the HVQVAE [16] model. The PSNR (dB) and the SSIM Fig. 10 Reconstructed face(d) using modified HVQVAE and (c) HVQVAE original version results on MaskedFace-CelebAHQ

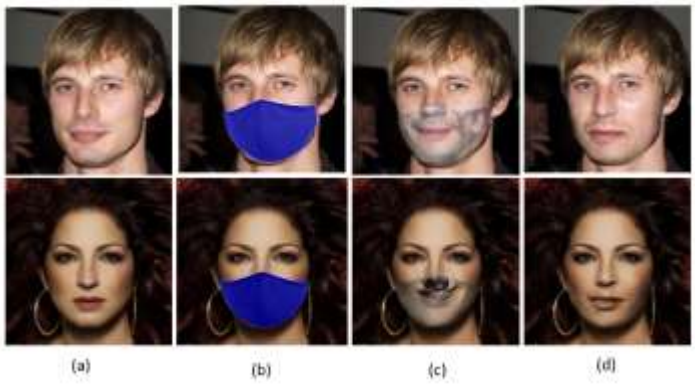

Fig. 10 Reconstructed face (d) using modified HVQVAE and (c) HVQVAE original version results on MaskedFace-CelebAHQ

numerical computations exactly show these results. In the case of the first ground truth image of the male subject, the PSNR and the SSIM values between (a) and (c) on the one hand and between (a) and (d) on the other are found to be 26.354 dB, 0.9148 and 29.423 dB, and 0.9558, respectively. For the second ground truth image of the female subject, these values are 28.127 dB, 0.9008 and 30.769 dB, 0.9478, respectively. As the modified HVQVAE model gave better results, further investigations will be based on the proposed HVQVAE model only, and three GAN-based techniques, EC, GC and modified HVQVAE, will be utilized for further investigation.

4.2.4 Comparative Assessment of the EC, GC and modified HVQVAE methods

After reviewing the existing literature and methods, we identified opportunities for improving face reconstruction from masked regions to obtain unmasked faces. We conducted simulations using the EdgeConnect [11], Gated Convolution [15], and Autoencoder-based HVQVAE [16] models on our MaskedFaceCelebA and MaskedFaceCelebA-HQ datasets. The EdgeConnect model was trained on the CelebA dataset, while the HVQVAE and Gated Convolution models were trained on the CelebA-HQ dataset; thus, all models were tested on both datasets. The EdgeConnect model performed well on MaskedFaceCelebA but poorly on MaskedFaceCelebA-HQ. The HVQVAE model showed good performance only on MaskedFaceCelebA-HQ, with no results on the MaskedFaceCelebA dataset. Similarly, the Gated Convolution model delivered superior results on MaskedFaceCelebA-HQ compared to MaskedFaceCelebA. These findings are presented in Tables 1 and 2 in Section 6. Our analysis reveals that the MaskedFaceCelebA-HQ dataset serves as the primary benchmark for evaluating and comparing the selected transfer learning models. As shown in Table 1, the Gated Convolution model performs comparably to the HVQVAE model and outperforms the EdgeConnect model on the MaskedFaceCelebA-HQ dataset.

Fig. 11 Qualitative Results 0n MaskedFace-CelebAHQ of testing of deep learning algorithms (a) Ground Truth (b) Edge Connect [11] (c) Gated Convolution [15] and (d) Modified HVQVAE [16]

EXPERIMENTAL RESULTS AND ANALYSIS

In our previous work [30], we investigated in depth the performance of learning- and non learning-based image inpainting techniques for reconstructing masked faces. Two state-of-the-art deep learning-based image inpainting models, namely, edgeconnect [11] and gated convolution [15], were examined and tested on our own two synthetic masked face datasets, namely, MaskedFace-CelebA and MaskedFace-CelebAHQ. These datasets are described in detail in [30]. The main focus when creating these datasets is to occlude faces with Covid-type masks. Both of these standard datasets are treated in the same fashion. These datasets are known as the MaskedFace-CelebA and MaskedFace-CelebAHQ datasets. Close analysis of the outcomes on the basis of the qualitative and quantitative results revealed that the MaskedFace-CelebAHQ dataset was better than the MaskedFace-CelebA dataset because it included high-quality images.

5.1 Qualitative analysis

The qualitative results refer to the image quality of the generated output in visual form and how close it looks to the original image. The comparative results are shown in Figure 11 and are based on the generated outputs for the MaskedFace-CelebA-HQ dataset images obtained using three deep learning models, namely, edgeconnect [1], gated convolution [15] and HVQVAE [16]. It is clear that the output of our proposed HVQVAE and gated convolution algorithm is better than the output produced by the edgeconnect model.

5.2 Quantitative analysis

The image quality assessment is carried out using the PSNR, SSIM, NCORR and UIQI metrics. Note that all four of these are full reference metrics. Among these, the SSIM and NCORR are perceived as better and more robust metrics by the research community. These image quality assessment metrics are summarized below:

5.2.1 Peak signal-to-noise ratio (PSNR): PSNR is a full-reference metric used to measure the signal-to-noise ratio in decibels, commonly applied to assess image and video quality after lossy compression. It requires identical dimensions for original and degraded images for accurate calculation. The formula for PSNR is:

$$\text{PSNR} = 20\ log_{10}\left(\frac{MAX_f}{\sqrt{MSE}}\right), \quad (1)$$

where MAX_f is the maximum signal value that exists in the original image and the mean squared error(MSE).

$$\text{MSE} = \frac{1}{mn} \sum_{0}^{m-1} \sum_{0}^{n-1} \|f(i,j) - g(i,j)\|^2 \ , \quad (2)$$

with f and g representing the original and degraded images, respectively, and m and n denoting the number of rows and columns.

5.2.2 Structural similarity measure (SSIM): SSIM is a full-reference metric for evaluating image quality based on structural properties. It is defined as:

$$SSIM(x,y) = \frac{(2\mu_x \mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)\ (\sigma_x^2 + \sigma_y^2 + c_2)} \quad (3)$$

where μ is the mean, σ is the standard deviation, x and y are the images compared, and c1, c2 are stability constants.

5.2.3 UIQI [34] evaluates image quality by considering luminance, contrast, and correlation. It is formulated as:

$$\text{UIQI} = (4 * \sigma_1 * \sigma_2 * \rho * \mu_1 * \mu_2)/(\sigma_1^2 + \sigma_2^2) * (\mu_1^2 + \mu_2^2) \quad (4)$$

where σ_1 and σ_2 are the standard deviations, ρ is the cross-covariance, and μ_1 and μ_2 are the means of the two images.

5.2.4 Normalized Cross-Correlation: Normalized Cross-Correlation measures image similarity by comparing the covariance of two images to their standard deviation:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]\ [n\sum y^2 - (\sum y)^2]}} \quad (5)$$

where x and y are pixel values. A value of 1 indicates a perfect match, 0 no correlation, and -1 a perfect negative correlation.

As mentioned earlier, we have carried out similar research in previous work[30], where we have concluded that better face reconstruction results are achieved using the Gated Convolution technique applied over MaskedFace-CelebAHQ. With further research, we now examine the same two datasets using hierarchical vector quantized variational autoencoders (HVAEs) [16], and a similar outcome is also observed in the present case. This is given in Table 1. Hence, our further investigations and observations will be based upon faces available in the MaskedFace-CelebAHQ dataset.

Table 1: Performance comparison of the HVQVAE[16] model on the MaskedFace-CelebA and MaskedFace-CelebAHQ datasets

Datasets	PSNR(dB)	SSIM	NCORR	UIQI
MaskedFace-CelebA	24.61	0.871	0.942	0.952
MaskedFace-CelebAHQ	25.97	0.898	0.979	0.977

Table 2 compiles the IQA outcomes for the four full-reference metrics for the three models we use in the present work. All these models, as described earlier, will operate upon the facial images of the MaskedFace-CelebAHQ dataset.

Table 2: IQA metrics for the three models on the MaskedFace-CelebAHQ dataset

Models	PSNR	SSIM	NCOR	UIQI
Hierarchical Vector Quantized Variational Autoencoders(HVQVAE)	25.97	0.898	0.979	0.977
Edge Connect	27.16	0.907	0.981	0.982
Gated Convolution	**28.12**	**0.921**	**0.988**	**0.989**

The following observations are based on Table 2. Table 2 for the image quality assessment metrics, PSNR, SSIM, NCORR and UIQI, clearly shows that the Gated Convolution method is better among the three facial reconstruction methods used in the present work—Edge Connect, Gated Convolution & HVQVAE. The numerical values of all four IQA metrics are the highest for Gated Convolution, while for Edge Connect, these are the lowest. Note that the original training of the Gated Convolution face reconstruction algorithm was performed on the CelebA-HQ facial image dataset. Similarly, the HVQVAE method is trained on the same dataset. In our proposed work, as mentioned earlier, we mildly tampered with the original CelebA and CelebA-HQ dataset images by occluding them with Covid-type masks instead of with normal rectangular occlusion. The edgeconnect face reconstruction scheme is trained on the CelebA dataset. Thus, compiling the computed numerical values for comparison yields that the MaskedFace-CelebAHQ dataset is the superior choice for carrying out further investigations. To compute the results presented in Table 2, therefore, only facial images from the MaskedFace-CelebAHQ dataset were used. Among the three models/schemes we examined in this work, the Gated Convolution model was found to be the most suitable for facial reconstruction purposes according to the compiled results in Table 2. It is crucial to note that maintaining a PSNR threshold of 36 dB is considered necessary to ensure that human observers cannot detect significant differences between the source and test images. A computed PSNR value less than 36 dB indicates noticeable visual dissimilarity and renders the images unsuitable for human assessment. As a result, alternative methods are required for subjective or quantitative evaluation of such images. The same principle applies to the SSIM, UIQI and NCOR evaluations.

It is important to emphasize that this study is in its preliminary stages, and further investigation is needed to validate the findings. Consequently, we subject these datasets to classification tasks to gather additional insights. By employing an efficient classifier, we can obtain quantified measurements of classification accuracy, as well as related parameters such as the F1 score, recall, and precision. A higher classification accuracy implies a lower degree of similarity between the subject images, as human observers equipped with the human visual system (HVS) struggle to differentiate highly similar sets of images. Hence, in the present context, achieving favorable PSNR, SSIM, UIQI and NCOR values would naturally result in reduced accuracy of the employed classifier. Further exploration of these aspects is addressed in Section 6.

Figure 11 shows the compilation of selected images from our MaskedFace-CelebA-HQ dataset obtained after the application of four different models, as mentioned above. The results compiled in Table 2 and Figure 11 corroborate this analysis. We call this part of our simulation PART-I.

Although it is ascertained in the PART-I dataset that the IQA results of our study produce better results for image quality assessment for the MaskedFace-CelebA-HQ dataset, the numerical values of the PSNR are well below the threshold value of 36 dB, which is critical for human perception. Therefore, a face recognition module certainly needs to be applied to the MaskedFace-CelebA-HQ dataset so that we can be doubly sure about the results compiled in

Table 2 and Figure 11. We carry out our extended simulation work only on the MaskedFace CelebA-HQ dataset, as it has given the best results according to Table 2.
The experimental work carried out for the face recognition task is organized as PART-II in our simulation. The results and analysis of PART-II are given in Section 6.

6. Recognition of Reconstructed Facial Images

In this section, we present our facial recognition results, focusing on the classification of faces into two categories: ground truth images and reconstructed images. To accomplish this, we employ a well-established ELM (extreme learning machine) classifier model. The model is trained using 80% of the available input data and evaluated on the remaining 20%. Now, let us provide a brief overview of the ELM classifier model used in our study, which facilitates the classification task by distinguishing between the ground truth images and the reconstructed image sets.

6.1 Extreme Learning Machine Model

The Extreme Learning Machine (ELM) is a single-layer feedforward neural network designed for fast training by randomly assigning input weights and biases and calculating output weights using the Moore-Penrose pseudoinverse [40-43]. Given a set of N training samples (x_i, y_i) where $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}^m$ and i = 1, 2, …, N. The ELM model uses $\hat{N}$ hidden neurons and an activation function g: $\mathbb{R} \rightarrow \mathbb{R}$. The output of the system [17] can then be given as:

$$\sum_{k=1}^{\hat{N}} \beta_k g(w_k x_i + b_k) = o_i \, \forall \, i \in 1, 2, \ldots, N \qquad (6)$$

Here, w_k is the weighting vector that connects the k^{th} hidden neuron to the input and output layers respectively, and b_k represents the threshold bias of the k^{th} hidden neurons. The hidden layer output matrix H is used to solve for β using the equation as given by Huang et al. [29], is:

$$\hat{\beta} = H^{\dagger} \qquad (7)$$

where $H^{\dagger}$ is the Moore–Penrose generalized inverse of the hidden–layer output matrix H[29]. Here, are brief descriptions of these metrics and their significance in evaluating the performance of the ELM classifier.

6.2 Accuracy: Accuracy measures a classifier model's performance by quantifying its ability to correctly identify both positive and negative samples. It is calculated as the ratio of correct predictions (true positives and true negatives) to the total number of samples:

$$\text{Accuracy} = \frac{(True\ Positive + True\ Negative)}{(True\ positive + False\ Positive + True\ Negative + False\ Negative)} \qquad (8)$$

6.3 Precision: Precision is crucial when the cost of false positives is high, as it evaluates the classifier's ability to correctly identify positive samples while minimizing the mislabeling of negatives as positives. It is calculated as the ratio of true positives to the total number of predicted positives:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad (9)$$

6.4 Recall: Recall, or sensitivity, measures the classifier's ability to correctly identify all actual positive samples (true positives). It is particularly important when the cost of false negatives is high, as it reflects the model's capacity for comprehensive detection of all positive instances. Mathematically, recall is calculated as:

$$\frac{True\ Positive}{True\ Positive+False\ Negative} \quad (10)$$

6.5 F1-Score: The F1-Score, also known as the F Score, is a metric used to evaluate the accuracy of a binary classification model. It combines the model's precision and recall, providing a balanced measure of performance. The F1-score ranges from 0 to 1, with 1 indicating perfect precision and recall and 0 indicating that either precision or recall is zero.

$$\text{F1 score} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

The accuracy, precision, recall, and F1 score collectively reflect the performance of a face recognition system, particularly in distinguishing reconstructed faces from ground truth faces. These metrics, along with the parameters used for IQA (image quality assessment) and mean absolute error (PSNR, SSIM, MAE, NC), are expected to exhibit an inverse relationship in the context of the experiments. Table 4 consolidates all the computed values generated by the ELM classifier.

6.6 Examining the Performance of the Classifier on Different Models for the MaksedFace-CelebA-HQ Dataset

To substantiate the understanding of the results compiled in Table 1 and Table 2, we carried out the following investigations using classification accuracy. We have two classes of images—the ground truth and the reconstructed images. Hence, any potential classifier should

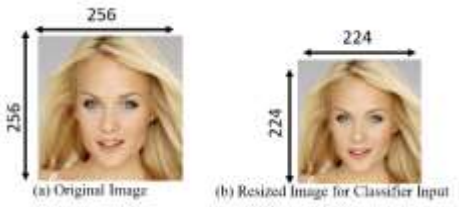

Fig 12: Depiction of resized image sample

be able to classify these images based on accuracy and other related metrics. In other words, if the reconstructed facial image is close to the ground truth counterpart, then the PSNR, SSIM, NCORR and UIQI will increase; otherwise, they will be less than the threshold values. Accordingly, in such a case, if the two sets of images quite resemble each other, then the classifier will not be able to properly classify the images and will eventually show poor classification accuracy. Thus, targeting classifier accuracy is a better way to determine and compare the results obtained after applying these facial reconstruction models. The images are resized to 224x224 for input to the classifiers Vgg16, Vgg19, ResNet50, ResNet101 and ResNet152 (Fig. 12).

6.7 Deep Transfer Learning Approach for Image Classification

In this case, we use a deep transfer learning approach for image classification with its basic classifier matched with the selected transfer learning model. For this purpose, Vgg16, Vgg19, ResNet50, ResNet101 and ResNet151 were used. We also attempt to replace the Softmax function with the Sigmoid function to evaluate the suitability of the classification task. The results involving both activation functions are compiled in Table 3 and Table 4, respectively, with and without fine tuning. These two tables compile only the classification accuracy. In the case of fine-tuning, only the end layer block is modified with the following configuration:

"conv5_block3_3_bn" and "conv5_block3_3_conv" for the ResNet50, ResNet101, ResNet152 models and

"conv5_block3_3_bn" for the Vgg16 and Vgg19 classifiers

All remaining blocks remained frozen. On the other hand, in the case of such a binary classification without fine tuning, all the blocks of the deep learning architecture are frozen.

Table 3: Performance of DCNN classifiers (without fine-tuning) on different models on the MaskedFace-CelebAHQ dataset

Model	Activation Fn	Vgg16	Vgg19	ResNet50	ResNet101	ResNet152	Avg
GC	Softmax	76.9	69	66.8	75	65.3	70.60
	Sigmoid	61.2	56.8	67.4	69	63.9	**63.66**
EC	Softmax	87.2	88.7	78.2	81.8	79.8	83.14
	Sigmoid	54.1	55.4	78.5	85.4	77.5	**70.18**
HVQVAE	Softmax	88.7	90.1	92.9	96.3	83.8	90.36
	Sigmoid	57	63.2	90.1	95.6	87.6	**78.70**

Table 4: Performance of DCNN classifiers (with fine-tuning) on different models on the MaskedFace-CelebAHQ dataset

Model	Activation Fn	Vgg16	Vgg19	ResNet50	ResNet101	ResNet152	Avg
GC	Softmax	98.9	87.9	70.4	81.3	80.3	83.76
	Sigmoid	89.7	73.8	67.2	76.3	74.8	**76.36**
EC	Softmax	99.7	96.5	81.9	78.3	79.1	87.10
	Sigmoid	83.8	81.1	87.9	87.7	80.2	**84.14**
HVQVAE	Softmax	93	93.6	92.2	92.1	92.1	92.6
	Sigmoid	91	87.3	94	92.8	89.5	**90.92**

Close observation of the results compiled in Tables 3 and 4 yields the following conclusions: Overall, the sigmoid activation function is better placed than the softmax activation function. This is because, for the sigmoid activation function, the classification accuracy is less than that for the softmax activation function. It is quite imperative that if the face reconstruction model is working fine, resulting in better face reconstruction, then as a result, the classifier will not be able to properly classify the original face from the reconstructed face (case of improved face reconstruction). In this scenario, all image quality assessment metrics give enhanced numerical values, while the classifier (operating with any activation function) provides reduced classifier accuracy. Hence, a deep insight into these computed and compiled values indicates that Gated Convolution yields the lowest average classification accuracy and the highest values for all the image quality assessment metrics. Thus, it can be concluded that GC is the best model for face reconstruction among the three selected models.

Table 5 compares the classification accuracies of the three face reconstruction models computed by using five different deep transfer learning schemes. The average classification accuracy of the algorithms is also compiled in this table. A close comparison is made on the basis of the average classification accuracy. It is very clear that in both cases without and with fine tuning the deep transfer learning architecture, the accuracy is the lowest for the Gated Convolution face reconstruction model. Thus, we safely conclude that Gated Convolution is the best face reconstruction model for the present research domain.

Table 5: Performance analysis of DCNN classifiers on different models on the basis of fine tuning and without fine tuning of classifier models. (Sigmoid Activation)

Model	Fine Tuning	Vgg16	Vgg19	ResNet50	ResNet101	ResNet152	Average
GC	without FT	61.2	56.8	67.4	69	63.9	**63.66**
	with FT	89.7	73.8	67.2	76.3	74.8	76.36
EC	without FT	54.1	55.4	78.5	85.4	77.5	**70.18**
	with FT	83.8	81.1	87.9	87.7	80.2	84.14
HVQVAE	without FT	57	63.2	90.1	95.6	87.6	**78.7**
	with FT	91	87.3	94	92.8	89.5	90.92

We have carried out Part II of the problem simulation in an organized manner. We used only the sigmoid activation function and not the softmax function, which is inferior to the sigmoid activation function. We used only the deep transfer learning-based standard classifier. In other words, we have not replaced the existing classifiers of these models with other potentially proven classifiers.

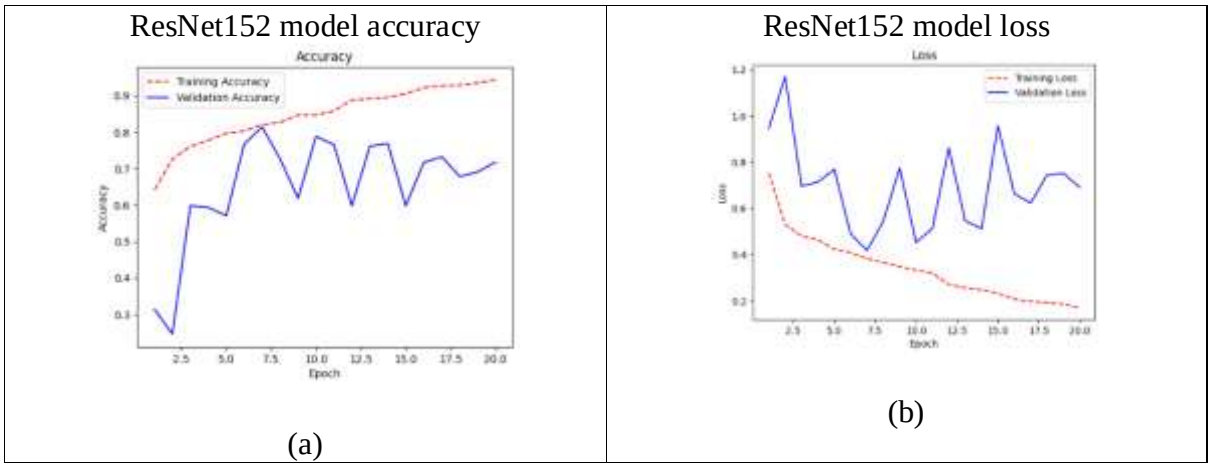

Fig. 13: ResNet152 Model (without Fine-tuning) for the MaskedFace-CelebAHQ dataset for the Gated Convolution Model

The training accuracy and validation accuracy are shown in Fig. 13 (a), whereas the training accuracy and loss are shown in Fig. 13 (b) in the form of a graphical representation of the ResNet152 classifier. In the last and third sections, PART-III, we plan to investigate the problem more intensively. We attempt to replace existing classifiers with several novel and established classifiers capable of producing better outcomes in different application domains, which are not necessarily linked with facial reconstruction purposes. These classifiers are extreme learning machine classifiers that have shown real-time classification capabilities[40]. Similarly, we also use the support vector machine (SVM) classifier, which is known for its robustness and generalizability[42]. However, this approach has not been proven to be as fast as other ELM-based classifiers. Thus, it has limited real-time capabilities[41].

Table 6: Performance analysis of classifiers trained with ELM and SVM (for the gated convolution model and sigmoid activation function without fine tuning) in terms of classification.

CNN-TL Models	Classifiers with accuracy (%)		
	Basic(Sigmoid)	SVM	ELM(sigmoid)
Vgg16	61.2	61.4	53.6
Vgg19	56.8	62.2	51.6
ResNet50	67.4	65.1	65.2
ResNet101	69	65.8	67.4
ResNet152	63.9	62.0	62.2
Average Accuracy	63.66	63.30	**60.09**

Table 6 compares the classification accuracy of the five considered deep learning transfer models in which the existing basic classifier is also replaced with a few established classifiers, as mentioned above. A close comparison of the results compiled in Table 6 shows that the classifier accuracy for all five deep transfer learning models followed by the ELM classifier, which presently works with the sigmoid activation function, is the minimum. The average accuracy is also found to be minimal. Among these five models appended with the SVM classifier, good results are also exhibited, more specifically for the ResNet deep transfer learning architecture. For the VGG deep transfer learning architectures Vgg16 and Vgg19, the basic classifier with the sigmoid activation function was found to be more suitable. Overall, the average classification accuracies of the basic classifier with the sigmoid activation function and that obtained from the SVM classifier are comparable. However, these two classifiers do not match the performance of the ELM classifier.

Therefore, for face reconstruction via the Gated Convolution approach, the ELM, which works as a single-layer feedforward neural network, is the best candidate for binary classification tasks. Moreover, face reconstruction using thousands of ground truth images is indeed a time-consuming task. The model performance, therefore, should be evaluated using a technique that is capable of exhibiting real-time classification outcomes. Therefore, the classifiers used in these five deep transfer learning models are henceforth used as ELM classifiers with sigmoid activation functions.

7. Ablation studies of the proposed work

Table 6 also provides additional insight. Among the five considered deep transfer learning models, the classification accuracy of ResNet152 (62.2%) was the closest to the average classification accuracy (60%). Therefore, we perform ablation studies for the following configuration of the ResNet152+ELM classifier with a sigmoid function. This ablation study will be performed by modifying and altering the values of different parameters and components. For this purpose, the following variations will be applied:

(i) Changing the number of hidden neurons in the ELM architecture.
(ii) The activation function is changed from sigmoid to three other activation functions, namely, Relu, leaky_ReLU, sigmoid and tanh.
(iii) The avg_pool layer was replaced with other layers.

7.1 Varying number of hidden neurons in the ELM architecture

Table 7 shows the training and testing accuracy (%) of the ResNet152+ELM classifier by varying the number of hidden neurons. The results show that the accuracy is lowest when the number of hidden neurons is 2048. However, we have shown that the accuracy of ResNet 152 + ELM is 62.2% with 64 hidden neurons.

Table 7: Varying the number of hidden neurons (S) in the ELM classifier with sigmoid activation

Number of hidden neurons	Training Accuracy(%)	Validation Accuracy(%)	Findings
16	80.6	63.5	Accuracy dropped
32	80.2	62.8	Accuracy dropped
64	82.7	62.2	Accuracy dropped
128	84.2	61.7	Accuracy dropped
256	84.9	61.4	Accuracy dropped
512	85.8	61.8	Accuracy dropped
1024	86.9	61.3	Accuracy dropped
2048	89.2	60.4	Least Accuracy
4096	92.7	60.4	Identical Accuracy

7.2 Replacing the Sigmoid Activation Function with Other Activation Functions

The ablation study shows that the ResNet152+ELM classifier yields the least accuracy for the model using the sigmoid activation function. Other activation functions, such as Tanh, Relu and Leaky_ReLU, are also tested, but the sigmoid function yields the best results, as shown in

Table 8: Changing the activation function from sigmoid to three other activation functions, namely, Relu, leaky_ReLU, Sigmoid and tanh (without fine-tuning), for the ResNet152+ELM classifier

Activation(ELM Classifier)	ResNet152+ELM classifier	Findings
Sigmoid	**60.9**	Least Accuracy
Tanh	62.3	Accuracy raised
Relu	61.2	Accuracy raised
Leaky_relu	61.1	Accuracy raised

7.3 Replacement of the avg_pool layer

Tables 9 and 10 show the four standard metrics obtained by replacing the avg_pool layer with other models and reveal that the global average 2D pooling algorithm yields the lowest accuracy for both the basic CNN-TL model and the ELM classifier.

Table 9: Replacement of the avg_pool (GlobalAveragePooling2D) layer in the ResNet152 model (basic DCNN-TL classifier)

Layer Name	Accuracy	Precision	Recall	F1_score	Finding
avg_pool	0.639	0.292	0.954	0.447	
GlobalMaxPooling2D	0.776	0.634	0.885	0.739	
Flatten layer	0.822	0.646	0.997	0.784	

Table 10: Replacement of the avg_pool (GlobalAveragePooling2D) layer in the ResNet152 model (basic CNN+ELM classifier)

Layer Name	Accuracy	Precision	Recall	F1_score	Finding
avg_pool	0.622	0.236	0.967	0.379	

GlobalMaxPooling2D	0.708	0.444	0.941	0.603	
Flatten layer	0.816	0.634	0.997	0.775	

Table 11: Results obtained from the five ablation studies.

Study No	Name of study	Best selected option obtained from ablation study	Accuracy(%)
1	Replacing the layer in ResNet152 model	avg_pool	62.3
2	Changing the classifier used along with ResNet152 model	ELM classifier	62.2
3	No. of hidden neurons in ELM classifier	2048 & above	60.4
4	Activation function used in classifier (S=2048)	Sigmoid	60.9
5	Changing the number of layers in ResNet model	152	62.2

Thus, a thorough ablation study carried out as described above clearly proves that the Gated Convolution (GC) is the best facial reconstruction scheme. This fact is critically evaluated by means of image quality assessment metrics, which use facial recognition classification accuracy under various parameters and circumstances and maintain the transfer learning architecture for a variety of parameter variations. The ablation (Table 11) proves Part I, Part II and Part III and clearly endorses these outcomes.

CONCLUSION

In conclusion, our research addresses the critical task of postmask face reconstruction, which is particularly significant amidst the challenges posed by the COVID-19 pandemic. We rigorously evaluated three prominent reconstruction algorithms—edgeconnect, gated convolution (GC), and hierarchical variational vector quantized autoencoders (HVQVAE)—using synthetic datasets, MaskedFace-CelebA and MaskedFace-CelebAHQ.

Our findings conclusively demonstrate that Gated Convolution (GC) outperforms Edge Connect and HVQVAE in terms of image quality and fidelity in face reconstruction tasks. Additionally, our study extended to face recognition tasks using a DCNN + ELM hybrid framework, where classifiers such as Vgg16, Vgg19, ResNet50, and ResNet101 performed excelled, with ResNet152 + ELM showing comparatively lower accuracy. This comprehensive analysis reaffirms GC as the preferred method for facial reconstruction, supported by its superior performance in both reconstruction and recognition applications. Our research contributes significant insights into overcoming contemporary challenges posed by widespread mask usage, advancing the field's understanding and application of facial reconstruction techniques.

Declarations

Availability of Data and Materials

- The CelebA and CelebA-HQ synthetic masked datasets generated during and/or analyzed during the current study are available in the **Synthetic masked dataset** repository (link: Google Drive). The folder includes the Masked Face dataset and the respective binary maps of masks for the CelebA and CelebA-HQ datasets in separate folders.

- The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request and were derived from the CelebA[5] and CelebA-HQ[14] public datasets.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

Not applicable.

Author Contribution Statement

- Chandni Agarwal: Conceptualization, Methodology, Software, Writing – review & editing.
- Charul Bhatnagar: Data curation, Visualization, Investigation, Validation.
- Anurag Mishra: Writing – Paper Organization, Experiment Organization and supervision.

Acknowledgments

Not applicable

REFERENCES

[1] Estudillo, Alejandro & Hills, Peter & Wong, Hoo Keat. (2021). The effect of face masks on forensic face matching: An individual differences study. Journal of Applied Research in Memory and Cognition. 10. 554-563.

[2] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2536-2544). https://arxiv.org/abs/1604.07379v2

[3] Y. Jiang, J. Xu, B. Yang, J. Xu and J. Zhu, "Image Inpainting Based on Generative Adversarial Networks," in IEEE Access, vol. 8, pp. 22884-22892, 2020, doi: 10.1109/ACCESS.2020.2970169.

[4] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In CVPR, pages 2536–2544, 2016.

[5] Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision (pp. 3730-3738). https://arxiv.org/abs/1411.7766v3

[6] Thung, K.-H., & Raveendran, P. (2009). A survey of image quality measures. 2009 International Conference for Technical Postgraduates (TECHPOS). doi:10.1109/techpos.2009.5412098

[7] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. UCTGAN: Diverse image inpainting based on unsupervised cross-space translation. In CVPR, pages 5741–5750, 2020.

[8] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoderdecoder with feature equalizations. In ECCV, pages 725–741, 2020.

[9] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In ACM SIGGRAPH, pages 417–424, 2000.

[10] Qin, Z., Zeng, Q., Zong, Y., & Xu, F. (2021). Image inpainting based on deep learning: A review. Displays, 69, 102028.

[11] Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z., & Ebrahimi, M. (2019). EdgeConnect: Generative image inpainting with adversarial edge learning. https://arxiv.org/abs/1901.00212v3
[12] Anwar, A., & Raychowdhury, A. (2020). Masked face recognition for secure authentication. arXiv preprint arXiv:2008.11104.
[13] Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2004, July). Extreme learning machine: a new learning scheme of feedforward neural networks. In 2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541) (Vol. 2, pp. 985-990).
[14] Karras et al., "Progressive Growing of GANs for Improved Quality, Stability, and Variation", in Internation Conference on Reoresentation Learning (ICLR), 2018
[15] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2019). Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4471-4480). https://arxiv.org/abs/1806.03589v2
[16] Peng, J., Liu, D., Xu, S., & Li, H. (2021). Generating Diverse Structure for Image Inpainting With Hierarchical VQ-VAE. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10775-10784)
[17] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In NIPS, pages 6306–6315, 2017.
[18] Xi Chen, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel. PixelSNAIL: An improved autoregressive generative model. In ICML, pages 864–872, 2018.
[19] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. arXiv preprint arXiv:1701.05517, 2017.
[20] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with PixelCNN decoders. In NIPS, pages 4790–4798, 2016.
[21] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In NeurIPS, pages 14866–14876, 2019.
[22] Yoshua Bengio, Nicholas Leonard, and Aaron Courville. ´ Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432, 2013.
[23] Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., & Li, H. (2017). High-resolution image inpainting using multiscale neural patch synthesis. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6721-6729). https://arxiv.org/abs/1611.09969v2
[24] Iizuka, S., Simo-Serra, E., & Ishikawa, H. (2017). Globally and locally consistent image completion. ACM Transactions on Graphics (ToG), 36(4), 1-14. https://doi.org/10.1145/3072959.3073659
[25] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, 23(10), 1499-1503.
[26] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., ... & Grundmann, M. (2019). Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172.
[27] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In CVPR, pages 1438–1447, 2019.
[28] Huang, G. B., Zhu, Q. Y., Siew, C. K.: Extreme Learning Machine: Theory and Applications. Neurocomputing, Elsevier, Vol. 70 (2006) 489 – 501.
[29] Huang, G. B., Zhu, Q. Y., Siew, C. K.: Real – Time Learning Capability of Neural Networks. IEEE Transactions on Neural Networks, Vol. 17, No. 4 (2006) 863 – 878.

[30] Agarwal, C., & Bhatnagar, C. (2023). Unmasking the potential: evaluating image inpainting techniques for masked face reconstruction. Multimedia Tools and Applications, 1-26.
[31] Agarwal, C., Itondia, P., & Mishra, A. (2023). A novel DCNN-ELM hybrid framework for face mask detection. Intelligent Systems with Applications, 17, 200175.
[32] Huang, G. B.: The MATLAB code for ELM. (2004) Available at: http://www.ntu.edu.sg/home/egbhuang
[33] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative image inpainting with contextual attention. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5505-5514). https://arxiv.org/abs/1801.07892v2
[34] Wang, Z., & Bovik, A. C. (2002). A universal image quality index. IEEE signal processing letters, 9(3), 81-84.
[35] Nizam Ud Din, Kamran Javed, Seho Bae, and Juneho Yi. 2020. Effective removal of user-selected foreground object from facial images using a novel GAN-based network. IEEE Access 8 (2020), 109648–109661.
[36] Nizam Ud Din, Kamran Javed, Seho Bae, and Juneho Yi. 2020. A novel GAN-based network for unmasking of masked face. IEEE Access 8 (2020), 44276–44287.
[37] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. 2021. Unmasking face embeddings by self-restrained triplet loss for accurate masked face recognition. arXiv preprint arXiv:2103.01716 (2021).
[38] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida, "Spectral Normalization for Generative Adversarial Networks," Proc. International Conference on Learning Representations (ICLR), 2018.
[39] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," Proc. Computer Vision and Pattern Recognition (CVPR), 21–26 Jul. 2017.
[40] Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2004, July). Extreme learning machine: a new learning scheme of feedforward neural networks. In 2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541) (Vol. 2, pp. 985-990).
[41] Agarwal, C., Mishra, A., Sharma, A., & Chetty, G. (2014). A novel scene based robust video watermarking scheme in DWT domain using extreme learning machine. Extreme Learning Machines 2013: Algorithms and Applications, 209-225.
[42] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.