IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 7, July 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007037 |

Facial Reconstruction of Masked Face using Generative Adversarial Networks (GAN)

Arpita Nanda, Dawood Damda, Gagan V, Emaad Jaffer, Madhusmita Mishra

UG Student, Dept. of I.S.E., The Oxford College of Engineering, Bangalore, Karnataka, India

UG Student, Dept. of I.S.E., The Oxford College of Engineering, Bangalore, Karnataka, India

UG Student, Dept. of I.S.E., The Oxford College of Engineering, Bangalore, Karnataka, India

UG Student, Dept. of I.S.E., The Oxford College of Engineering, Bangalore, Karnataka, India

Assistant Professor, Dept. of I.S.E., The Oxford College of Engineering, Bangalore, Karnataka, India

ABSTRACT: There have been significant advances in facial recognition software and surveillance technology. With these advancements came changes to the systems to overcome other issues of facial recognition software like pose variation and expression variation. However, facial occlusions like masks have become increasingly prevalent as a safety measure but also as a means of avoiding detection by facial recognition and surveillance software. This has become a major cause of failures in these systems. To combat these failures, a system has been developed that reconstructs the obstructed facial features of the individual using Generative Adversarial Networks (GAN). A Generative Adversarial Network is a deep learning based generative model that works by using two components called a generator and discriminator that compete against each other to become more precise with their predictions. The areas where facial features are obstructed are first mapped and isolated during the image processing phase and then the isolated areas are reconstructed by the generator. To obtain realistic images that are close to ground truth, the algorithm is trained extensively with a diverse dataset. For coarse-to-fine image completion, the system runs through a higher number of training epochs. The project aims to support law enforcement agencies with surveillance and help with investigations where the suspects facial features are occluded by a mask or similar object. This system can help create a positive identification of the suspect and therefore help reduce crime in society.

KEYWORDS:Facial Recognition, Surveillance, Occlusion, Mask, Reconstruct, Facial features, Generative Adversarial Networks, Generator, Discriminator, Prediction, Mapped, Isolated, Epochs, Identification.

I. INTRODUCTION

Facial recognition softwares and surveillance technology have been evolving exponentially since their induction as they find extensive application in improving security. These systems have faced multiple issues like pose variation, variation in illumination, variation in expression, ageing, similar faces, image resolution and last but not the least, occlusions [1]. Most of these issues have been satisfactorily resolved using different techniques like the Principal Component Analysis (PCA) method for pose variation [2] and the Contractive Convolutional Network forvariation in expression [3]. So, this system focuses on occlusions as this issue is a challenging one to overcome.

The most common gear used by fringe elements of society while committing any crime is a facial covering of some sort that obscures part of their facial features, generally the nose, mouth and sometimes the forehead. The parts that remain are the eyes and the areas surrounding it. The main aim of using such facial covering is to avoid detection and identification either by individuals or any facial recognition and surveillance technology.

The system aims to use only the data of the eyes and the areas surrounding it toreconstruct the entire face of the individual in order to produce an image which can be run through facial recognition software and to generate IDs on suspects. This in turn will help law-enforcement agencies and aid their investigations of crime.

The system will use Generative Adversarial Networks to reconstruct the obstructed part of the masked face. This will allow for accurate reconstruction using Convolutional Neural Network (CNN).

The proposed system is a model that has the ability to make clear distinction between the face and the mask by being able to segment these areas. The system then goes on to map out this area and following that, generate the reconstruction of the areas below the masked region be making use of the data from around the masked region.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007037 |

II. RELATED WORK

Literature survey of the proposed system revealed several methods where Generative Adversarial Networks have been used in the context of human faces and their manipulation. The survey also revealed that most of the work done on previous system were not performed with facial occlusions as their primary focus but rather on other factors.

Yujun Shen et al. [4] proposed a system to change the angle at which a face is presented while preserving the facial features. They used a three-player GAN unlike the usual two-player GAN to change the angle of the face from a side profile to a frontal profile. The images are generated by making use of the natural symmetry that faces have. The image is first laterally inverted and then stitched on to the original image to present a face that is now facing the user.

Shuang Liu et al. [5] presented a method to edit the attributes of faces by changing the latent variable by making use of a pre-trained conditional GAN and a linear classification model. The process that this system uses is differentiated into two stages: First, depending on the optimization function, the generative model performs a latent variable search to construct a high-quality image that is similar to the input image. Then, a latent variable of the network is edited by which the features of the constructed image can be modified indirectly. This means that the process is not affected by the training process and the network structure of the GAN. This system can prove to be incredibly helpful to generate sketches of individuals when only a handful characters and features are made available. The generated images can then be used to perform further investigation and analysis.

Nizam Ud Din et al. [6] presented a user-friendly system to remove obstructions from facial images where the user of the system has the control over which object is to be removed. They system can be used as many numbers of times as required by the user and is shown to be quite effective through the result analysis. The effectiveness of the system is showcased using a few commonly occurring obstructions in images. The objects are hands, microphones, sunglasses and eyeglasses. The model is trained to identify a user-selected object in the first level. In the second level, the object is removed by using the object detection information that was identified in the first level. The second level of the system specifically involves the integration of both partial and vanilla convolution operations in the generator part of the network. The network is trained using a paired synthetic face occluded dataset. The model is evaluated using real world images from the Internet and the CelebA dataset.

Xiang Chen [7]et al. have proposed a system that reconstructs the entire face of an individual by making use of the data from around their eyes. This system focuses on the eye region of the face to reconstruct the rest of the face while aiming to preserve the original facial features. The system is an end-to-end network based on conditional generative adversarial networks and is developed to generate the facial information based only on the available data of the eyes region. To generate images that preserve information and are accurate, close to ground truth, a synthesis loss function based on feature loss, GAN loss and total variation loss is proposed to guide the training process. This system only focuses on generating images from the data made available from the eye region. It does not focus on any other occlusions and any occlusions on the eye region can be problematic.

As noticed from the extensive literature survey performed, a lot of systems use Generative Adversarial Networks (GAN) for applications involving manipulating and generating facial features. The popular technology colloquially and commonly called as deepfakes also makes use of Generative Adversarial Networks as part of their method to generate realistic depictions of individuals in situations that the individual may have never been in. This is an up-and-coming technology that is gaining traction as we progress.

From the literature survey conducted, a need for a system which focuses on all types of occlusions was noticed. So, the proposed system is a method is which the entire face region is taken as input, the obstruction is then analysed and the system can generate a reconstruction based on wherever the obstruction is noticed. The system does not require any user intervention to detect and remove the occlusion which makes it desirable for use in various scenarios. The literature survey also made clear the need for a diverse and varied dataset as most systems made use of a generic dataset consisting mostly of Caucasian individuals which made the system inefficient and inaccurate when reconstructing or manipulating the facial features of individuals from a more diverse set.

The inclusion of the above features i.e., the ability to use the entirety of the face region, the implementation of a diverse dataset and the fact that the proposed system is completely automated with no real requirement of user intervention during the reconstruction process makes the system more desirable. The addition of these features will mean that the proposed system is more dynamic for use in various use cases including the uses involving law enforcement agencies and other intelligence entities.

III. PROPOSED METHODOLOGY

Facial coverings like masks have become increasingly prevalent all across the world following the COVID-19 pandemic. This brought forward a complex issue for facial recognition systems as they are not able to function as

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007037 |

intended. The proposed system is presented to solve these issues and aid the process of facial recognition systems. This will aid the investigative process of law enforcement agencies and help keep the society safe.

The goal of facial reconstruction of masked face using Generative Adversarial Networks (GAN) is to produce accurate and realistic images that are close to ground truth so that these images can be used for real world applications like running them through facial recognition software. This is performed by a set of convolution neural networks that work independently to execute different phases of the reconstruction. The system architecture can be broken down into a few simple tasks as shown below in figure 1.

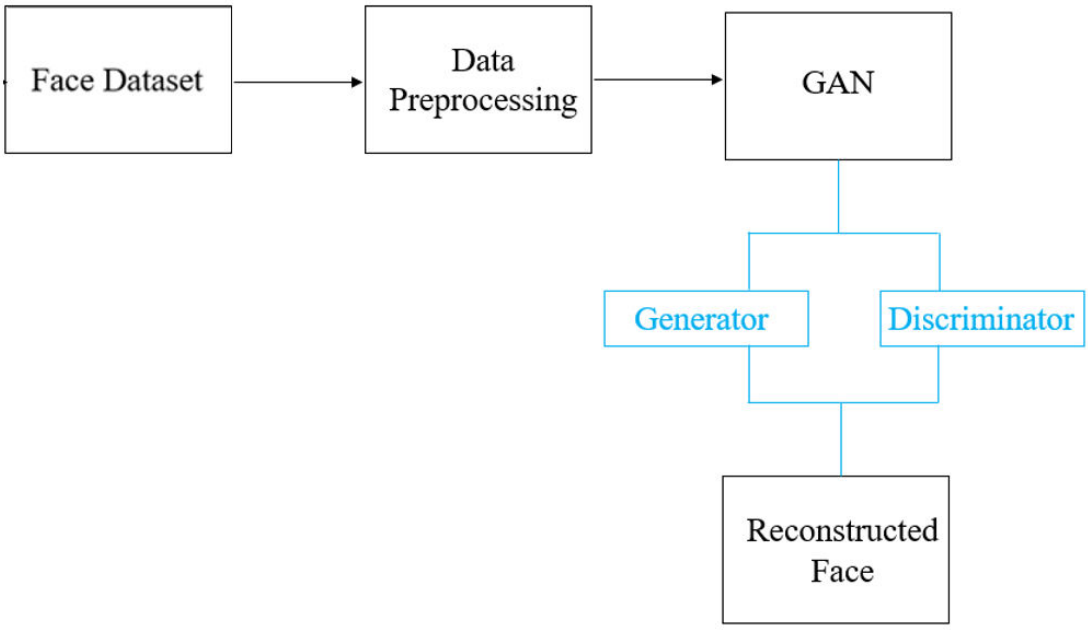

Fig. 1: A simplified system architecture diagram of the proposed method

A. *Description of Dataset:*

A dataset [8] consisting of 4,980 images of faces, called Labeled Faces in the Wild (LFW) is taken.The dataset is then augmented to resize them= images to 128x128. Since datasets of faces with masks are not readily available, images of masks are superimposed onto these images to create a dataset that consists of masked faces. Three types of commonly available masks have been used: surgical masks, cloth masks and the N95 masks. These masks have been chosen as they are the most prevalent all across the world. This results in two datasets that are available for the model to use.The two datasets (masked and not masked) are split and used for different aspects of training and validation.

B: *Creating maps of the areas over which the mask is present:*

In order to find out the part of the face that is to be reconstructed using GAN, the area over which the mask is present needs to be isolated first. During the training phase, this is done using the cv2 module of the OpenCV library in python. The masked and unmasked images of the same individual are superimposedto find the masked area and an image of the mapped area is created and saved. The absdiff() function in OpenCV finds out the parts of the image which are varying and creates an image of the varying parts. This function finds the values in an image that are an absolute difference of the original unmasked image. This results in an extremely accurate map of the area over which the mask is present. This process yields another dataset consisting of the maps of all the masks in the images. An example of the maps created is shown in figure 2.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007037 |

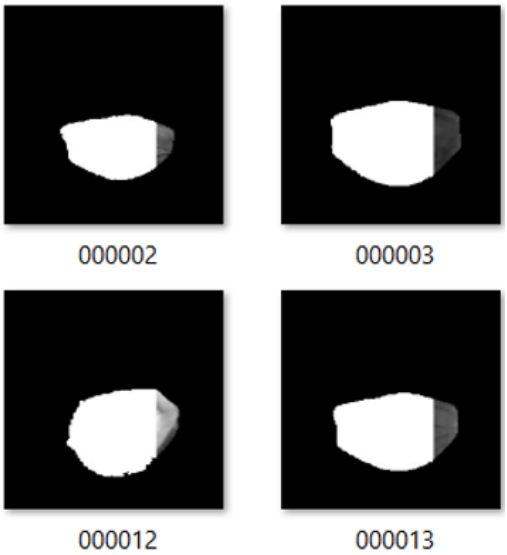

Fig. 2: Maps of masked areas

C: *Generator part of Generative Adversarial Networks:*

Generator model of generative adversarial network checks what part of the face is obscured and generates the image for it. The accuracy of thereconstruction improves as the number of training epochs are increased. The maps generated in the processing phase of the model are used to identify the area over which reconstruction (generation) is supposed to take place. The generator performs the reconstruction in the mapped area and superimposes it on to the masked face. The reconstruction takes the place of the mask in the image. This image needs to be appealing both visually and in the manner that it should be close to ground truth.

Unlike the usual U-Net implementation in the generator model, the input image will not have the masked image (128x128 single channel) but the same 128x128, three-channel image that is linked with the generated segmentation map. The output will be a 128x128 three-channel image. Subsequently, it is also noticed that the introduction of a block named Squeeze and Excitation [9]

D: *Discriminator part of Generative Adversarial Networks:*

The discriminators used in this model present five convolutional levels that have batch normalization and leaky ReLU. The only exceptions are the in the first level which lacks batch normalization and the last level which lacks both batch normalization and Leaky ReLU. This model of Generative Adversarial Networks has two discriminators for specific purposes. The difference between he inputs of these discriminators is that the whole region discriminator takes the image created by the generator or ground truth image (128x128 in three channels) while the mask region discriminator takes the input from the generator.

The mask region discriminator checks for inconsistencies only in the reconstructed part that was created by the generator. The discriminator will either pass it off as not good enough or the generator will have been advanced enough to circumvent the discriminator. This discriminator does not check the surrounding region but rather just focuses on the generated area that was performed by the generator part of the module. If the reconstruction seems accurate enough without a lot of noise and with features that are complete, without any incomplete areas, the mask region discriminator will validate this output. The discriminator gets stronger with every epoch that the system runs, so the higher the number training epochs, the more accurate the system gets.

The whole region discriminator checks to see whether the face as a whole looks consistent with the reconstructed part. This discriminator will not allow any shoddy reconstructions to get past it. This includes the generated are being properly aligned with the rest of the face. The whole region discriminator will make sure the reconstruction and the rest of the face are seamless with their borders and the entire image as a whole is good enough to now be used for real world applications.

E: *Perceptual Network:*

The final part of the editing module is the perceptual network. The perceptual network is implemented to generate the facial images which have features that are as close to ground truth as possible. This can be achieved by the use of a pre-trained VGG19 on ImageNet [10]. In this model, the feature map of three convolutional levels (block3_conv4,

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007037 |

block4_conv4, block5_conv4)have been exploited. It receives three channel 128x128 images and images of ground truth as input

F: *Accounting for various losses:*

While the module is being trained the network needs to learn the high-level features of the face such as the nose, mouth, chin, cheekbones so that it can create the perfect reconstruction. To achieve this, in addition to the generic GAN loss, others were also used to refine the image and reduce the noises in the generated image and also increase the accuracy of the generated image. A custom loss function called reconstruction loss is defined which is represented as follows:

$$\mathcal{L}_{rc} = \mathcal{L}_{SSIM} + \mathcal{L}_{l1}$$

The loss of the generator is a non-saturating loss [11]:

$$\mathcal{L}_{adv} = \mathbb{E}_{I_{edit} \in S}\left[\log\left(D\left(G\left(I_{input}, I_{mask\_map}\right)\right)\right)\right]$$

The goal of introducing this loss function is to maximise the probability that the discriminator labels the generation as true. This is achieved by maximising the implementation of this loss on both the discriminators. The loss of perceptual network is based, as mentioned above on the feature maps of the intermediate levels of VGG19.

All these losses are then combined to allow the generator to create accurate reconstructions while learning all the high-level features of the face.

$$\mathcal{L}_{comp} = \lambda_{rc}\left(\mathcal{L}_{rc} + \mathcal{L}_{perc}\right) + \lambda_{whole}\left(\mathcal{L}_{D}^{whole} + \mathcal{L}_{adv}^{whole}\right) + \lambda_{mask}\left(\mathcal{L}_{D}^{mask} + \mathcal{L}_{adv}^{mask}\right)$$

Where the parameters $\boldsymbol{\lambda_{RC}}$,$\boldsymbol{\lambda_{whole}}$ and $\boldsymbol{\lambda_{mask}}$have the values 100, 0.3 and 0.7 respectively.

The losses of both discriminators of minmax type:

$$\mathcal{L}_{D}^{whole} = \mathbb{E}_{I_{gt} \in O}\left[\log\left(D_{whole}\left(I_{edit}, I_{gt}\right)\right)\right] + \mathbb{E}_{I_{edit} \in S}\left[\log\left(1 - D_{whole}\left(G\left(I_{input}, I_{mask\_map}\right)\right)\right)\right]$$

$$\mathcal{L}_{D}^{mask} = \mathbb{E}_{I_{gt} \in O}\left[\log\left(D_{mask}\left(I_{mask\_region}, I_{gt}\right)\right)\right] + \mathbb{E}_{I_{edit} \in S}\left[\log\left(1 - D_{mask}\left(G\left(I_{input}, I_{mask\_map}\right)\right)\right)\right]$$

Where $\boldsymbol{I_{gt}}$ is the ground truth image.

IV. EXPERIMENT

The operation of the two modules i.e., the mad module and the editing module was implemented separately. The training of the editing module (GAN) proved to be more complicated due to the high complexity of the networks involved. The network was trained for forty epochs two gain satisfactory results that were visually appealing and also looked close to ground truth. The model accuracy and loss values have been plotted as shown below in figure 3.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007037 |

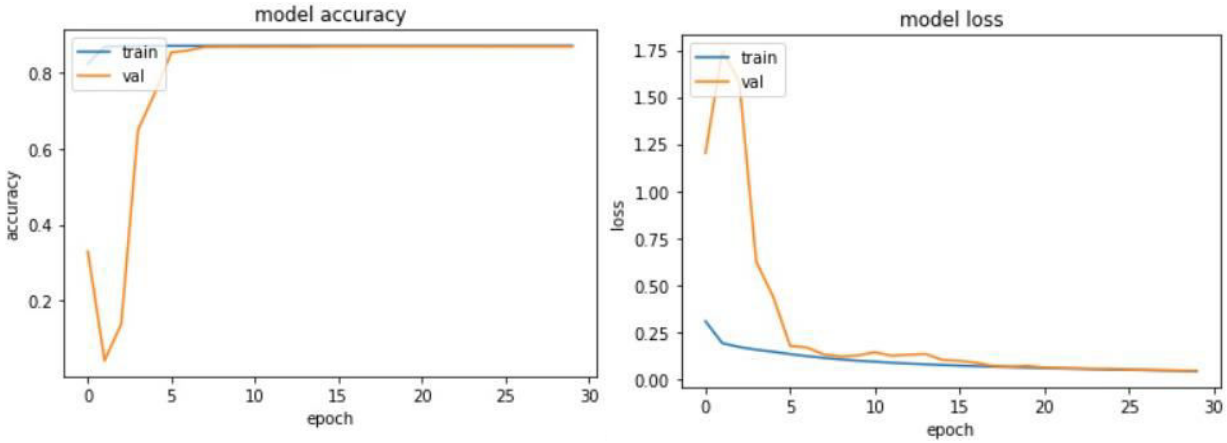

Fig. 3: Trends in accuracy and loss during the training phase

From the data shown in the graphs, it is observed that the values of accuracy are quite low at first but rise exponentially as the number of epochs increase showing that the model aand losses have been successfully implemented. The same is also noticed for the values of losses, as the training phase progresses in terms of number of epochs, the loss values exponentially fall to nearly approach zero. These were the results of experiments performed on the training of the map module.

For the other part of training i.e., the training of the editing module, the network was initially reduced to a three block network as the computaional power required for the traditional system would be too high. The activation function was also replaced with a sigmoid function. The result of this alteration proved to be undesirable as the resultant images were washed out, pixelated with flattened colours as observed in figure 4.

Fig. 4: The observed output in three block network

This problem was identified as the normalization fapplied to the input images not being consistent with the sgmoid that was being used as activation function. This was solved by changing the output trigger function and through the replacement with a tanh. (with value between -1 to 1) Through this change, the results were observed to be much better as shown in figure 5.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007037 |

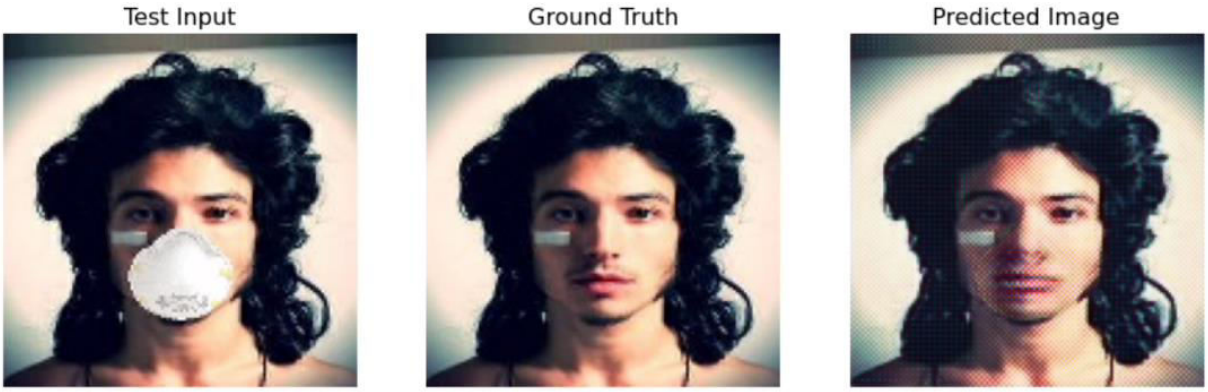

Fig. 5: Output after replacement with tanh

Subsequently, an attempt was made to increase the number of blocks from three to five but this proved to be fatal for the computer as it resulted in a ResultExhaustedError as the system ran out of GPU and RAM space. After analysis, a problem was identified in the squeeze and excitation block. By rectifying this problem, the model was able to work successfully with five bocks in the network.

V. RESULTS

The results obtained from the system using the Labeled Faces in the Wild have proven to be suitable and up to the standard. This evaluation is made not only by visually observing the generated images but also through appropriately selected image metrics. Two metrics have been used for the evaluation. The first metric is Structural Similarity Index Measure (SSIM) which measures the similarity between two images. This is a full reference metric [12] which means that the measurement of image quality is based on an initial uncompressed or distortion free image as reference. The second metric is Peak Signal-to-Noise Ratio which is a ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation [13]. PSNR is expressed using a logarithmic quantity using the decibel scale.

Satisfactory results of SSIM and PSNR have been observed throughout the functioning of the system with a highest Structural Similarity Index Measure of 0.5 and a high Peak Signal-to-Noise Ratio of nearly 20dB. These results are shown below in figure 6.

Fig. 6: Results of value metrics obtained from the system

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165

|| Volume 10, Issue 7, July 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1007037 |

VI. CONCLUSION AND FUTURE WORK

As observed in the above sectons from the value metrics, imporovements can certainly be made in various areas to get better scores of structural similarity index measure and peak signal to noise ratio. These improvements can be made by increasing or decreasing the number of parameters to make the dataset more heterogenous. The architecture of the model can be further improved, number of parameters can be modified and network complexity can be changed. These changes can be made within the system. A real improvement however, can be brought about by rendering a dataset that is much more heterogenous which can be difficult in terms of complexity and time involved since this would mean that collection of a dataset with thousands of images.

Further improvements to this system can be made in term of simplicity to use by providing a simple and robust user interface which would allow the end user to input the images with mask and the system would return a satisfactory result of the facial image without the mask. This system can also be made more useful and easy to use by interfacing it with a facial recognition software. This addition to the software will make it more useful and powerful in temrs of capability to identify a higher number of individuals with a lesser chance of failure due to obstructions and occlusions on the facial image.

A subsequent improvement to this system would be to perform the operation of unmasking the face in real time which would prove to be incredibly helpful and quich and speedy identification. This would prove to be extremely advantageous in terms of aiding the investigation process of law-enforcement agencies.

REFERENCES

[1] “Challenges in Face Recognition Systems”, Merrin Mary Solomon, Mahendra Singh Meena, Jagandeep Kaur, Electronics and Communication Engineering Department, Amity University, Haryana, India

[2] “Face Recognition Performance in Facing Pose Variation”, Alexander A. S. Gunawan and Reza A. Prasetyo, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

[3] “Disentangling Factors of Variation for Facial Expression Recognition”, Salah Rifai, YoshuaBengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. Universit´e de Montr´eal Department of Computer Science and Operations Research

[4] Y. Shen, P. Luo, P. Luo, J. Yan, X. Wang and X. Tang, "FaceID-GAN: Learning a Symmetry Three-Player GAN for Identity-Preserving Face Synthesis," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 821-830, doi: 10.1109/CVPR.2018.00092.

[5] S. Liu, D. Li, T. Cao, Y. Sun, Y. Hu and J. Ji, "GAN-Based Face Attribute Editing," in IEEE Access, vol. 8, pp. 34854-34867, 2020, doi: 10.1109/ACCESS.2020.2974043.

[6] N. U. Din, K. Javed, S. Bae and J. Yi, "Effective Removal of User-Selected Foreground Object From Facial Images Using a Novel GAN-Based Network," in IEEE Access, vol. 8, pp. 109648-109661, 2020, doi: 10.1109/ACCESS.2020.3001649.

[7] X. Chen, L. Qing, X. He, J. Su and Y. Peng, "From Eyes to Face Synthesis: a New Approach for Human-Centered Smart Surveillance," in IEEE Access, vol. 6, pp. 14567-14575, 2018, doi: 10.1109/ACCESS.2018.2803787.

[8] “LFW Face Database : Main.” LFW Face Database : Main, http://vis-www.cs.umass.edu/lfw/.

[9] J. Hu, L. Shen, S. Albanie, G. Sun e E. Wu, "Squeeze-and-Excitation Networks" 2017.

[10] Krizhevsky, I. Sutskever e G. Hinton, "ImageNet Classification with Deep Convolutional" 2012.

[11] M. Lucic, K. Kurach, M. Michalski, S. Gelly e O. Bousquet, "Are GANs Created Equal? A LargeScale Study" 2017.

[12] Wikimedia Foundation. (2021, December 26). Peak signal-to-noise ratio. Wikipedia. From https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio

[13] Wikimedia Foundation. (2022, July 14). Structural similarity. Wikipedia. From https://en.wikipedia.org/wiki/Structural_similarity

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details