A GAN-Based Approach for Unmasking masked faces

***Abstract*— This project creates a robust model capable of analyzing masked facial features and generating realistic predictions of the underlying faces with applications in security and forensics. Our main goal is to remove mask objects on the face in facial images and generate accurate predictions of the underlying faces. Our project utilizes two main modules: the map module for detecting face masks and generating binary segmentation maps, and the editing module for removing masks and reconstructing the uncovered regions using a modified U-Net with two discriminators. Additionally, we improve our model with a GAN-based network featuring two discriminators—one for understanding the overall face structure and another for focusing on elaborating the missing details. This enables our model to effectively remove masks and generate realistic facial features. The project's outcome will be predictive images of masked individuals, revealing their underlying facial features. These images could be useful in identity verification and forensic investigations, aiding in the process of identifying individuals who are wearing masks.**

Index terms: Map module, Editing module, Binary Segmentation maps, Modified U-Net, Generative adversarial network.

1. Introduction

The need of wearing masks has grown rapidly in recent years, driven by concerns about pollution, diseases and a desire for privacy. But with this taken as consideration, there has been challenges in scenarios of facial recognition. Removing masks, which covers almost half of the face, might be helpful in identifying/revealing individual's identity, which would be helpful in different situations. Our work will focus on finding a solution for object removal, and focusing on unmasking a masked face.

Our research focuses on removing mask object from facial images and reconstruct the face according to the trained model from CelebA dataset. We propose a novel GAN-based approach with two discriminators that detects and removes masks, and completing the uncovered facial regions. Previous methods for object removal relied on patch matching, limiting their effect on large objects like masks.

Our approach use new techniques in machine learning which excel at object removal but face challenges with complex objects like masks. In order to address this, we divide the task into two phases: the detection of mask objects and the completion of the identified mask region's image. In the first stage, we detect the mask object from facial images and generate binary segmentation map. In the second stage, we use one generator and 2 discriminator GAN setup. One discriminator looks at the whole image and enforces global coherency. While the other discriminator focuses on deep missing semantics that looks only at the missing region, ensuring realistic completion of the face.

To train our model, we constructed a matched synthetic dataset by altering photos from the publicly accessible CelebA dataset, as there are no facial image pairs with and without mask objects.

2. Related Work

I. Paper 1

This papers uses an Expression-Conditioned Generative Adversarial Network (ECGAN) which is designed for reconstructing the masked faces with particular expressions.

The ECGAN uses binary segmentation maps of masks and expressions to produce high-quality images, giving great results in generating realistic images with particular expressions. RAFDB dataset is used for training and testing the model, which is used to generate realistic images with particular expressions with great effectiveness, exceeding baseline and state-of-the-art approaches. This approach combines mask segmentation using R-CNN, a generator, and two discriminators. It generates more realistic image and accurate expressions. The result will suggest that with reconstruction faces with having facial expressions can be applicable in various situations.

The paper helps us in our project with GANs. The GAN's se of binary segmentation maps of face masks and expressions to produce high-quality facial images will be useful in mask removal and facial feature reconstruction. This could help recreate faces look more realistic and accurate.

3. Proposed Approach

Our research focuses on reconstructing a realistic and accurate facial image using a Generative Adversarial Network (GAN) for accurate identification and recognition of individuals who are wearing masks or other facial covering.

It consists of two main modules, map module and editing module.

A) MAP MODULE

The map module provides a binary segmentation map called Imask. 1 represents the mask object, while 0 represents the remaining pixels in the image. Gmask, a CNN-based map generator, consists of an encoder and decoder that use a modified U-Net architecture. Except for the first layer, the encoder section of the generator is made up of five convolution blocks, each of which represents a convolution layer followed by a Leaky ReLU activation function and an instance normalisation layer. The decoder is a replica of the encoder, except instead of a convolution layer, it utilises a deconvolution layer. The final layer of the decoder employs the tanh activation function without a normalising layer. Additionally, by concatenating the output of the deconvolution layers with feature maps from the encoder at the same level, skip connections are employed to mix local and global information. The generator uses an encoder network to down-sample an input image to the bottleneck layer, then it up-samples the image to forecast a binary map. There is a cross-entropy loss between the target and projected maps.

B) EDITING MODULE

This module's objective is to take off the mask and finish the face picture's missing portion in a way that makes it consistent with the ground truth image. It consists of a perceptual network, discriminators, and editing generators.

1) EDITING GENERATOR

The Editing Generator, Gedit, shares the same design as the Map Generator. However, it includes some additional capabilities for removing the mask and filling in the missing areas of the face. The generator concatenates the input picture, Iinput, with the map module, Imask_map, to produce a generated image, Iedit_gen.

Iedit_gen = Gedit_gen(Iinput, Imask_map).

The editing generator utilizes a combination of two forms of loss: structural similarity loss (SSIM) and L1 loss to drive the output image to look realistic.

LRC = Ll1 + Lssim

L1 loss is used to measure the difference in pixels between generated image Iedit_gen and the ground truth (real image) Igt.

Ll1 = |Iedit_gen – Igt|.

SSIM is used to measure the structural similarity between Iedit and Igt.

Lssim = 1 – SSIM(Iedit_gen, Igt).

2) DISCRIMINATORS

There are two discriminators used, Dwhole and Dmask_region. The Dwhole approach helps in ensuring that the output created by the generator maintains the arrangement of the original input image.

$$L^{whole} = -E_{Igt \in O} \log Dwhole(Iedit\_gen, Igt) + E_{Iedit\_gen \in S} \log(1 – Dwhole \times (Gedit\_gen(Iinput, Imask\_map)))$$

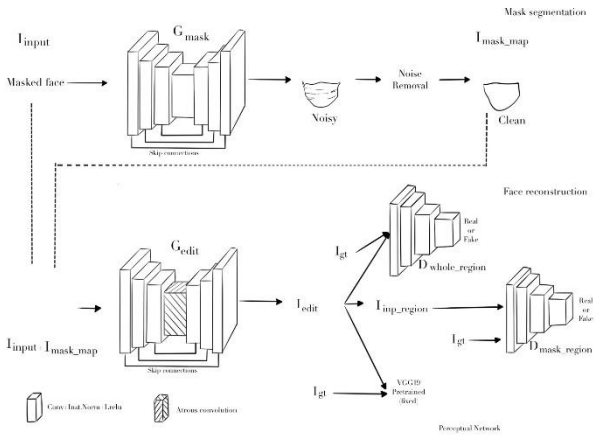
Here, O and S are the original and synthesized image sets, respectively.

The Dmask approach is used in missing region only. This approach helps to generate realistic facial image in the mask removal region.

$$L^{mask} = -E_{Igt \in O} \log Dmask(Imask, Igt) + E_{Iedit\_gen \in S} \log(1 – Dmask \times (Gedit\_gen(Iinput, Imask\_map)))$$

Here,

Imask = Iinput × (1 – Imask) + (Iedit_gen × Imask)



For training in GAN, generator minimizes the loss functions:

$L^{whole} = - E_{Iedit\_gen \in S} \log(Dwhole \times (Gedit\_gen(Iinput, Imask\_map)))$

$L^{mask} = - E_{Iedit\_gen \in S} \log(Dmask \times (Gedit\_gen(Iinput, Imask\_map)))$

3) Perceptual Network

The Perceptual Network is a pre-trained VGG-19 network. It encourages the generator's output to have comparable characteristics to the ground truth image. It helps generator's accuracy by making the output look more like the real image. It does this by instructing the generator to pay more attention in important details and features.

A perceptual loss function, Lperc, is used. It compares the features of Iedit and Igt.

Let ψi is the activation map of the i^{th} layer of ψ, the perceptual loss is defined as:

$$Lperc = \sum_i ||\psi(Iedit\_gen) - \psi(Igt)||$$

A joint loss function is defined which combines all the loss functions, to train the editing module:

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

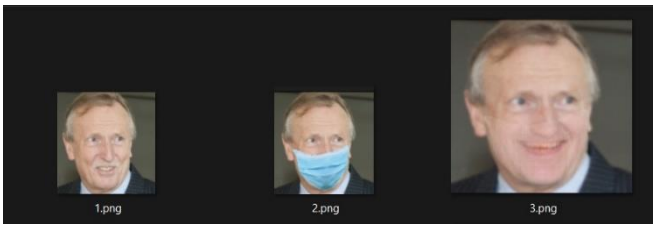
- Binary Segmentation U-Net

Binary Segmentation U-Net separates the mask region on the face in facial image from the remaining uncovered area. It divides the facial image into sections, where one section of the image reduces the size to effectively capture its details. And the other section increases the image size to reconstruct these details accurately. A standardized U-Net is used with additional techniques like dropout and instance normalization to improve stability and model's performance.

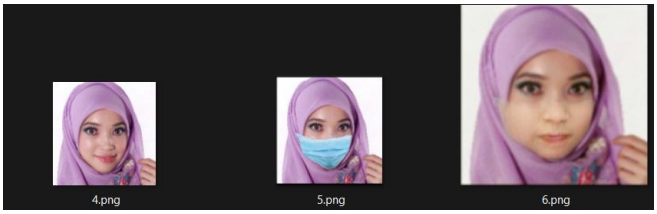
Loss Functions:

- **Adversarial Loss:** $L_{adv} = \min_G \max_D \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$
- **Reconstruction Loss:** $L_{recon} = ||y - G(x)||_1$
- **Hybrid Loss:** Combines adversarial loss with L1 loss for the generator to ensure both realistic and accurate inpainting.

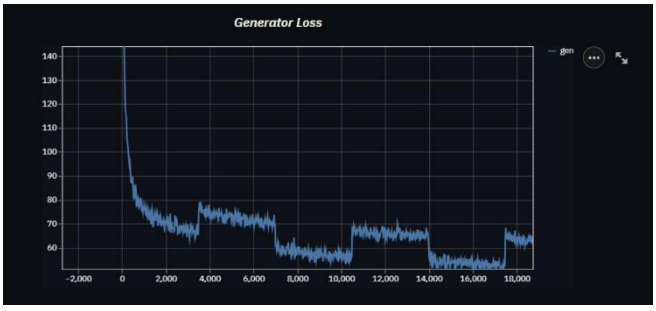
4. Results



The first image in the above sample shows the ground truth image of an old man. The second image shows the masked face of the same old man which covers most of his face. The third image shows the predicted image of the same old man with removing the mask and predicting the covered region by our project.



The first image in the above sample shows the ground truth image of a woman. The second image shows the masked face of the same woman which covers most of her face. The third image shows the predicted image of the same woman with removing the mask and predicting the covered region by our project.



This is a graph for Generator Loss.

The accuracy is **0.72**.

5. Conclusion and future work

I. Conclusion

Our project shows the approach our team used for reconstructing the face obstructed by a face mask. Our project first segments the image to determine the part of the face where the mask exists and then reconstructs this part of the original image. We are able to see that the model generates an approximation of what the person looks like. However we notice that the image generated is not totally flawless, we can see places where the reconstructed face is blurry and also places where the skin color doesn't fully match the target image. We have realized that this is likely due to the less number of epochs that were executed while the model was learning. We have learnt that fine tuning of the model parameter along with higher epoch numbers will create a model with higher accuracy. Our model contributes to the ongoing research on face reconstruction and highlights the impact of DL approaches for the same.

I. Future Work

- Refine the model learning further to improve the accuracy of the image segmentation and the reconstruction of the faces to match the target image more.
- Improve the speed of giving the output and implement mechanisms to allow real time reconstruction of faces from CCTV footage and video streams.
- Train the model on a more diverse dataset in order for the model to be capable of reconstructing the faces of people from various ethnicities, races and age.
- Extend the model to recognise and reconstruct faces hidden by other object such as scarfs, caps and helmets.