A UNET++ and CoGAN-based method to remove face masks from the masked faces

Akhil Kumar  
SCSET, Bennett University

Divyam Gupta  
CSE, UIET, Panjab University

Manisha Kaushal  
Thapar Institute of Engineering & Technology

Akashdeep Sharma (akashdeep@pu.ac.in)  
CSE, UIET, Panjab University

Research Article

Keywords: CoGAN, Pix2Pix, UNET++, PatchGAN, image to image translation

Posted Date: September 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3351025/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Image-to-image translation emerges as a significant utility of conditional Generative Adversarial Networks (CoGANs). This research introduces a fresh application of conditional GANs, aiming to uncover hidden facial attributes. Our methodology involves enhancing the Pix2Pix GAN framework through the integration of a modified UNET++ architecture, which serves as an inventive generator model. In this setup, the Pix2Pix model employs a PatchGAN architecture within the discriminator, producing an activation map with values utilized to authenticate depicted faces. Through the incorporation of the UNET++ architecture into the generator, we effectively narrow down the semantic gap between the encoder and decoder feature maps. This strategic adjustment results in a noticeable enhancement in gradient flow. To gauge the effectiveness of our proposed approach, we conducted experiments on a bespoke dataset intentionally crafted for training paired image-to-image translation GANs. Our model is comprehensively compared against other leading models designed for revealing concealed facial features. Significantly, our proposed model convincingly surpasses these alternatives across a range of evaluation criteria.

1. Introduction

Generative Adversarial Networks (GANs) [1] constitute a deep learning approach to generative modeling, often utilizing convolutional neural networks. The GAN framework encompasses a pair of neural networks, specifically the Generator and the Discriminator, engaged in a competitive dynamic to create novel images. The Generator produces batches of samples, juxtaposed with genuine domain instances, and feeds them to the Discriminator for classification into real or synthetic categories. This adversarial interplay reflects a zero-sum game within game theory. The prowess of Generative Adversarial Networks (GANs) in producing remarkably authentic and finely detailed images has fueled their exceptional performance in image generation. Among the array of successful image translation applications for GANs, such as CoGAN [2], cycleGAN [3], stackGAN [4], and more, this study centers on harnessing the Pix2Pix [5] network for its research objectives. However, a constraint of the GAN model lies in its potential to generate arbitrary domain images. The conditional GAN, or CoGAN, counteracts this by enabling targeted image generation of specific categories. This involves introducing an additional input layer in the form of one-hot-encoded image labels, guiding the generator during training to synthesize images of a designated type. In the context of face mask removal using GANs, the process entails two steps: (1) identifying the mask, generally by delineating the mask region's boundary, and (2) generating the concealed region obscured by the mask. These steps collectively contribute to the final synthesized image.

The present study puts forth a variant of the Pix2Pix GAN to facilitate the generation of unmasked facial images from obscured and blurred inputs, employing a pair-to-pair image translation paradigm. Furthermore, we introduce a meticulously curated dataset that establishes a direct correspondence between input and output images, embodying the transformation from masked to unmasked representations. The principal contributions of this paper encompass:
1) We introduce a novel GAN architecture designed to unveil concealed facial features. In pursuit of effective image-to-image translation, we present an innovative adaptation of the CoGAN-based Pix2Pix GAN framework. Our approach involves the integration of the UNET ++ architecture into the generator model, alongside a customized array of convolutional networks for the discriminator models. The outcomes of our proposed model are noteworthy, with SSIM values of 0.91115 and 0.85752, as well as PSNR values of 31.7251 and 24.9883, observed on the training and test sets, respectively. These results are indicative of its ability to uphold both structural fidelity and signal clarity amidst noise when reconstructing the facial region obscured by a mask.

2) Furthermore, we introduce a dataset comprising depictions of individuals donning masks juxtaposed with corresponding images of the same individuals without masks. This comprehensive dataset encompasses more than 30,000 facial images, capturing diverse instances of individuals wearing three distinct types of masks.

3) Comparative analyses reveal the notably superior performance and heightened suitability of the proposed GAN within real-world contexts. When contrasted with analogous approaches, our GAN model achieved a significant enhancement, showcasing an improvement ranging from 1.2–5.9% in SSIM scores and a further advancement of 2.82–6.35% in PSNR metric scores.

This work encompasses the subsequent sections: In Section 2, an exploration of the pertinent literature within the realms of image editing and face mask removal is presented. Section 3 delves into the materials and methodologies employed, elucidating the formulation of the dataset and the development of the model. The experimental assessments, outcomes, and limitations of this endeavor are expounded upon in Section 4. Finally, Section 5 encapsulates the conclusion drawn from this study and outlines prospects for future advancements within the domain.

2. Related work

GANs have established their presence across diverse domains, encompassing tasks such as object manipulation and generation, image and sound synthesis, and high-resolution image creation. This section provides an overview of GAN advancements with a specific focus on image generation, particularly concerning concealed portions of images and facial structures.


Examining related research on object removal, image editing, and face mask removal and reconstruction underscores the scarcity of techniques proficient in identifying face masks, removing them, and accurately reconstructing the unmasked image. Addressing these gaps, our work endeavours to devise a method capable of adeptly removing face masks, facilitating accurate synthesis of unmasked facial images.

3. Materials and methods

3.1 Dataset

The proposed approach employs a technique of paired image-to-image translation, necessitating a specialized dataset that ensures a one-to-one correspondence between input and output images. As such, the creation of a tailored dataset featuring these two distinct image categories was imperative. The input images depict individuals with masks covering their faces, while the target images portray the same individuals without masks, representing their unobstructed facial appearance.

Given the limited availability of suitable publicly accessible datasets, our work took the initiative to curate a customized dataset utilizing images of renowned Bollywood celebrities, conveniently accessible on Kaggle [20]. This dataset encompasses a comprehensive collection of 12,355 images showcasing the top 100 Bollywood personalities. To enhance the dataset's realism, we also incorporated sample images from the dataset used by authors in [21], thus augmenting the dataset's diversity and mitigating the risk of neural networks converging to local minima. Figure 1 offers a glimpse of the proposed dataset's structure. The dataset preparation involved cropping facial regions and eliminating redundant images, resulting in a final count of 11,262 images without face masks. To introduce variety, the dataset featured three distinct types of face masks superimposed on the images: white face masks, Anti-COVID face masks, and 3M face masks. Consequently, the modified dataset encompassed 11,262 images of
unmasked faces, as well as equal numbers of images for each type of mask. It's important to note that the method employed for mask removal from facial regions draws upon both the images of unmasked faces and faces wearing masks.

Incorporating OpenCV's DNN module [24], we proficiently render face masks onto facial regions. This module aids in extracting facial landmarks, generating a comprehensive set of 68 landmarks across the face, encompassing regions like eyes, eyebrows, nose, mouth, and jawline. These landmarks subsequently serve as reference points for precisely overlaying face masks as needed. To ensure optimal mask alignment, we thoughtfully annotated mask points corresponding to key landmarks along the jawline's base and the top of the nose. The alignment process employs the OpenCV library to determine transformations between matched key points, mask annotations, and corresponding facial landmarks. Subsequently, the identified transformation matrix is applied to map these points accurately. To bolster the proposed model's proficiency in predicting face mask regions, we undertook the training of a tiny YOLO v4-SPP face mask detector [22]. The face mask detector utilized in this study was employed to identify areas with face masks in a publicly available face mask detection dataset [23]. This dataset encompasses a substantial compilation of 52,635 images featuring individuals both with and without face masks. The regions identified by the tiny YOLO v4-SPP face mask detector were extracted, resulting in cropped images of the detected face mask areas within the images that depicted mask-wearing. These identified face mask regions, denoted by black rectangular boxes, were subsequently overlaid onto the modified Bollywood Celebrity dataset. By strategically integrating these highlighted mask regions into the altered images, we facilitated the precise identification of various types and complexities of face masks by the proposed model. This streamlined the process of removing masks from obscured faces. The comprehensive process of mask area detection via the tiny YOLO v4-SPP face mask detector is elucidated in Fig. 2, while the intricate steps of dataset creation are meticulously elaborated upon in Fig. 3.

3.2 Approach

In this section, we provide an in-depth insight into the proposed GAN architecture, which is founded upon a modified UNET++ encoder-decoder design for the purpose of unveiling concealed facial features. The operational intricacies of this architecture are expounded upon. Figure 4 visually represents the comprehensive architecture adopted within this study. The underlying model takes an input image and proceeds to identify the presence of a facial mask. Subsequently, it undertakes the task of reconstructing the obscured facial region that lies behind the detected mask. The subsequent sections delve into the architecture's specifics and mechanics, unraveling its inner workings in detail.

a) Discriminator module

The discriminator's role involves processing both a source image (a facial image with a mask) and a target image (a facial image without a mask). It then assesses the likelihood that the target image is either a genuine or a synthesized transformation of the source image. The architecture of the
discriminator is intricately tied to the effective receptive field of the model, which determines the ratio between the model's output and the number of input image pixels. For this endeavor, the PatchGAN architecture was chosen for the discriminator. This design meticulously aligns each output prediction of the model to a 70x70 square patch within the input image. This strategy offers the advantage of adaptable applicability to input images of varying sizes, whether larger or smaller than 256x256 pixels. The model's output can take the form of a singular value or a square activation map, featuring values that assess the authenticity of each patch in the input image—whether it's real or fake. These values can be averaged to provide an overall probability or classification score if desired. The PatchGAN's structure is influenced by the effective receptive field size, often referred to as the receptive field. This term denotes the relationship between a model's output activation and a region within the input image. Notably, the receptive field isn't equivalent to the discriminator model's output size. Instead, it pertains to a model that outputs a pixel in the activation map corresponding to a pixel in the input image. Traditionally, the receptive field is defined based on the size of an individual convolutional layer's activation map relative to its input, filter size, and stride value. The 70x70 PatchGAN adheres to a fixed number of three layers (excluding the output and second-to-last layers), regardless of the input image's dimensions. The formula for calculating the one-dimensional receptive field is: 

$$ \text{receptive field} = (\text{output size} - 1) \times \text{stride} + \text{kernel size} $$

where output size represents the previous layer's activation map size, stride signifies the filter's application shift, and kernel size indicates the filter's dimensions. Furthermore, the discriminator module processes two concatenated input images and generates a prediction matrix, as depicted in Fig. 5. It consists of a series of layers including Convolution, Batch Normalization, and LeakyReLU. The proposed discriminator module encompasses six Convolution layers, with a kernel size of 4 and a stride of 2 for the initial four layers. The training of the discriminator employs the Adam optimizer with a learning rate of 0.0001, a beta_1 hyperparameter of 0.5, and the Poisson loss function. The model comprises a total of 6,968,257 parameters (2,816 trainable and 6,965,441 non-trainable parameters). The model's weights are initialized with a normal distribution characterized by a standard deviation of 0.02.

b) Generator module

The generator module within the proposed CoGAN-based Pix2Pix architecture adopts the UNET architecture, characterized by its distinctive "U" shape. This symmetrical structure comprises two integral parts: the contracting path, situated on the left, and the expansive path forming the right side of the U. The contracting path employs convolution processes, while the expansive path employs 2D transpose convolution layers. This encoder-decoder model employs skip connections, linking equivalent layers between the encoder and decoder blocks that share matching feature map sizes.

In this work, we incorporate a modification of the UNET architecture known as UNET++. UNET ++ refines the traditional skip connections present in UNET by addressing the semantic gap between encoder and decoder feature maps before concatenation. This architecture utilizes nested and dense skip connections, effectively capturing intricate details of 2D images.
The UNET++ structure comprises encoder and decoder blocks interconnected through a sequence of nested dense convolution blocks, as depicted in Fig. 6. It differs from the original UNET architecture in several key ways:

1) It integrates convolution layers within skip pathways, narrowing the semantic gap between encoder and decoder feature maps.

2) It introduces dense skip connections on skip pathways to enhance gradient flow.

3) It incorporates deep supervision, enabling model pruning and performance enhancement.

The UNET++ architecture enhances segmentation accuracy by incorporating Dense blocks and Convolution layers between the encoder and decoder. It achieves this by redesigning the skip pathways to bridge the semantic gap between encoder and decoder sub-paths. These convolution layers aim to minimize the semantic gap between the feature maps of the encoder and decoder sub-networks, simplifying optimization for the optimizer.

Dense skip connections, inspired by DenseNet, are implemented within UNET++ as skip pathways between the encoder and decoder. These Dense blocks accumulate previous feature maps and deliver them to the current node through the dense convolution block along each skip pathway. This generates high-resolution feature maps at multiple semantic levels.

Capitalizing on UNET++'s advantages, we propose a modified version of the original Pix2Pix GAN, featuring a generator module built on the UNET++ architecture. The generator module comprises an encoder sub-network or backbone followed by a decoder sub-network. Each encoder block is composed of Convolution layers, BatchNormalization layers, and ReLU activation layers. The encoder network consists of five such blocks, all with a kernel size and stride of 2 and identical padding. The decoder block's upsampling layers employ Convolution transpose layers with a kernel size and stride of 2 and similar padding. The final output layer employs a convolution layer with a kernel size of 1 and a Tanh activation function. This choice of activation ensures pixel values in the generated image lie within the range of [-1, 1]. The generator module is updated using a weighted combination of Poisson and Mean Squared Error losses, with the weighting skewed toward Mean Squared Error. The Adam Optimizer is employed to minimize the losses, featuring a learning rate of 0.0001 and beta_1 of 0.5.

To propose a method for unmasking masked faces, we integrate the previously described discriminator and generator modules, creating the UNET++ and CoGAN-based Pix2Pix GAN model. This combined GAN model operates with a substantial 23,023,172 parameters, out of which 16,053,603 are trainable, and 6,969,569 are non-trainable.
c) Loss function

The GAN model we propose employs a combination of Poisson and Mean Squared Error losses to address issues of vanishing gradients and convergence. The specifics of these two loss functions are outlined as follows.

1) Poisson loss

In the context of our proposed GAN model, the Poisson loss function serves as a tool for regression, particularly suited for modeling count data. This loss function gauges the dissimilarity between the projected outcome and the observed actual output. Mathematically, the Poisson loss can be expressed through Eq. (1).

\[
L(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N} \left( \hat{y}_i - y_i \log \hat{y}_i \right)
\]

The act of minimizing the Poisson loss is synonymous with maximizing the likelihood of the data, assuming that the target follows a Poisson distribution given the input. This loss is applicable when there's a belief that the target value adheres to a Poisson distribution and the aim is to model the rate parameter based on a particular input.

2) Mean Squared Error

The Mean Squared Error (MSE) serves as a widely used loss function in regression scenarios. This loss function quantifies the average of squared disparities between predicted and actual values across the provided data. MSE is particularly responsive to outliers; when confronted with numerous instances featuring identical input features, the optimal prediction is their mean target value. This stands in contrast to the Mean Absolute Error, where the median becomes the optimal prediction. MSE is especially suitable when the target data, contingent on the input, adheres to a normal distribution centered around a mean value. Furthermore, it's beneficial when the intention is to penalize outliers effectively. The mathematical representation of the Mean Squared Error (MSE) loss is captured by Eq. (2).

\[
L(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N} (y_i - \hat{y}_i)^2
\]

In Eq. (1) for Poisson loss and (2) for Mean Squared Error (MSE) loss, \( y \) represents the actual output whereas, \( \hat{y} \) is the predicted expected value for \( i \) data count and \( N \) computations.
The previously outlined generator and discriminator, coupled with the described loss functions, exhibited a favorable loss curve during training. This curve reflects successful training on the training set, contributing to the accomplishment of the image-to-image translation objective. Graphical representations of the generator loss, discriminator loss, and the comprehensive GAN model loss are depicted in Fig. 7.

4. Experiments and Results

In this section, we delve into the results garnered from a series of experiments conducted to validate the efficacy of the proposed approach. The UNET ++ and CoGAN-based Pix2Pix model, designed for unmasking masked faces, was implemented on a system featuring the following configuration: Intel® Core i7-6850K CPU @ 3.60 GHz x 12, coupled with 64 GB of RAM, and two NVIDIA TITAN XP 12 GB GPUs. The dataset was partitioned into training and testing subsets, maintaining a 55:45 ratio. The training subset, constituting 55% of the images, was employed for generating realistic images of unmasked faces using the generator module. To facilitate the discrimination between generator module outputs and real images of unmasked faces, the discriminator module was presented with outputs from the generator module along with a collection of 17,603 genuine unmasked face images. For evaluation purposes, a distinct set of 14,000 images from the testing subset, exclusive of the training set’s masked face images, was utilized. The training phase of the proposed approach spanned 300 iterations, employing a batch size of 4. The training process was facilitated by the Adam optimizer, leveraging a learning rate of 0.0001 and a beta_1 value of 0.5. In the generator module of the proposed GAN model, the Mean Squared Error (MSE) loss and Poisson loss were utilized, while the discriminator module relied solely on the Poisson loss. Further elaboration includes insight into the evaluation metrics, the sequence of experiments, the resultant evaluation outcomes, and a comparative analysis against other relevant methodologies. These aspects are meticulously presented in the subsequent subsections.

4.1 Evaluation metrics

The central emphasis of this investigation centers on image-to-image translation, making the evaluation of structural similarity between images and their signal-to-noise ratio of utmost importance. To quantify the structural similarity, the SSIM (Structural Similarity Index) metric has been employed. Concurrently, the PSNR (Peak Signal-to-Noise Ratio) metric has been adopted for evaluating signal-to-noise ratio. The SSIM metric, denoting the Structural Similarity Index, is a robust tool for gauging the quality of digital images and videos. Its purpose lies in assessing the likeness between two images, utilizing criteria such as luminance intensity, contrast, and structural characteristics. This metric encapsulates these factors to provide an estimate of similarity between the images. Mathematically, the SSIM metric is articulated through Eq. (3).

$$SSIM (x, y) = l(x, y) * c(x, y) * s(x, y)$$
In Eq. (3), the variable "l" denotes the luminance, which is used to compare the brightness of two images. The variable "c" represents contrast, indicating the variation in intensity range between the brightest and darkest regions of the two images. Meanwhile, "s" stands for structure, enabling a comparison of the local luminance between the two images to ascertain their similarity. The coordinates \((x, y)\) symbolize the two images being analyzed.

Another metric employed in this study, the PSNR (Peak Signal-to-Noise Ratio), is indicative of the highest achievable signal-to-noise ratio. It fulfills the role of quantifying quality assessment in situations involving quality degradation due to various codecs and image compression methods. The PSNR metric involves treating the original image as the signal and the distortion caused by image compression as the disparity. The computation of PSNR involves determining the Mean Squared Error (MSE) – a measure of the squared differences between a noise-free monochrome image and its noise-affected approximation. The mathematical expression for Mean Squared Error (MSE) in the context of a given noise-free image and its noise approximation is elucidated by Eq. (4).

\[
MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2
\]

In Eq. (4), \(m\) and \(n\) are dimensions of a monochrome image \(I\) and \(K\) is the value of noise approximation.

The PSNR can be expressed using Eq. (5).

\[
PSNR = 10 \cdot \log_{10} \left( \frac{MAX^2_I}{MSE} \right)
\]

In Eq. (5), \(MAX^2_I\) is the maximum possible pixel value of the image.

4.2 Evaluation results

The method we put forth underwent training on an approximate count of 17,000 images encompassing both masked and unmasked faces, subsequently undergoing testing on nearly 14,000 images. The training process was conducted over the span of 300 epochs, employing a batch size of 4. This choice of a smaller batch size was deliberate, aimed at averting potential out-of-memory issues. To assess the effectiveness of the proposed GAN model, its performance was gauged using evaluation metrics including SSIM and PSNR, as detailed in Table 1. This assessment was carried out on both the training and test datasets, thereby encompassing images akin to those in the training set and images distinct from it present in the test set.
Table 1
Metrics findings for the proposed GAN model

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SSIM</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set</td>
<td>0.91115</td>
<td>31.7251</td>
</tr>
<tr>
<td>Testing Set</td>
<td>0.85752</td>
<td>24.9883</td>
</tr>
</tbody>
</table>

As illustrated in Table 1, the GAN model introduced in this study achieved notable results with an SSIM value of 0.91115 and a PSNR value of 31.7251 on the training set. Similarly, on the test set, the model yielded an SSIM value of 0.85752 and a PSNR value of 24.9883. These outcomes underscore the model's capability to reconstruct the concealed facial regions behind masks, both on similar and diverse images, demonstrating commendable structural similarity and adeptly preserving signal integrity amidst noise. To qualitatively assess the model's performance, we subjected select images from the training set to scrutiny, evaluating its efficacy in reconstructing the obscured facial regions. These reconstructions were then juxtaposed against the ground truth. The results from this qualitative assessment, as depicted in Fig. 8, were promising and augured well for the model's performance.

4.3 Ablation experiments

The outcomes derived from the proposed GAN model showcased remarkable results in both quantitative and qualitative aspects. However, for the sake of validating the model's performance, we undertook an exploration of five additional combinations, each involving alterations in the number of layers within the generator and discriminator modules. These explorations encompassed diverse configurations of encoder and decoder layers, resulting in variations in the tally of trainable and non-trainable parameters. Nonetheless, the empirical investigations established that the most favorable results, as measured by SSIM and PSNR, were consistently attained with the initially proposed GAN model configuration. This specific configuration entailed a generator module comprised of 4 encoder layers, 1 bottleneck layer, and 10 decoder layers, alongside a discriminator module equipped with 6 layers. Detailed insights into the configuration of these custom experiment setups can be gleaned from Table 2. Further quantified assessments of performance, relying on SSIM and PSNR metrics, can be observed in Tables 3 and 4. In addition, Fig. 9 offers a visual representation of the qualitative outcomes derived from employing various model configurations. These experimental endeavors provide valuable insights into the judicious selection of generator and discriminator layers, a decision-making process pertinent not only to the specific challenge of face mask removal addressed in this study, but also applicable to broader image-to-image translation tasks.
### Table 2
Description of the custom GANs trained for comparison

<table>
<thead>
<tr>
<th>Model</th>
<th>Generator Layers</th>
<th>Discriminator Layers</th>
<th>Trainable Parameters</th>
<th>Non-Trainable Parameters</th>
<th>Total Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>6 encoder, 1 bottleneck, 6 decoder</td>
<td>6</td>
<td>67,097,603</td>
<td>11,177,473</td>
<td>78,275,076</td>
</tr>
<tr>
<td>Model 2</td>
<td>7 encoder, 1 bottleneck, 7 decoder</td>
<td>6</td>
<td>54,423,299</td>
<td>11,179,329</td>
<td>65,602,628</td>
</tr>
<tr>
<td>Model 3</td>
<td>12 encoder, 1 bottleneck, 12 decoder</td>
<td>6</td>
<td>41,250,691</td>
<td>11,179,073</td>
<td>52,429,764</td>
</tr>
<tr>
<td>Model 4</td>
<td>43 encoder, 1 bottleneck, 43 decoder</td>
<td>14</td>
<td>48,885,315</td>
<td>19,314,433</td>
<td>68,199,748</td>
</tr>
<tr>
<td>Model 5</td>
<td>7 encoder, 1 bottleneck, 7 decoder</td>
<td>6</td>
<td>82,736,899</td>
<td>11,180,353</td>
<td>93,917,252</td>
</tr>
<tr>
<td>Proposed</td>
<td>4 encoder, 1 bottleneck, 10 decoder</td>
<td>6</td>
<td>16,053,603</td>
<td>6,969,569</td>
<td>23,023,172</td>
</tr>
</tbody>
</table>

### Table 3
Evaluation results of custom GANs on training set

<table>
<thead>
<tr>
<th>Model</th>
<th>SSIM</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>0.85640</td>
<td>28.4696</td>
</tr>
<tr>
<td>Model 2</td>
<td>0.88068</td>
<td>30.1880</td>
</tr>
<tr>
<td>Model 3</td>
<td>0.75869</td>
<td>23.8032</td>
</tr>
<tr>
<td>Model 4</td>
<td>0.76087</td>
<td>22.2019</td>
</tr>
<tr>
<td>Model 5</td>
<td>0.87019</td>
<td>29.4226</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.91115</td>
<td>31.7251</td>
</tr>
</tbody>
</table>
Table 4
Evaluation results of custom GANs on test set

<table>
<thead>
<tr>
<th>Model</th>
<th>SSIM</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>0.83119</td>
<td>25.8888</td>
</tr>
<tr>
<td>Model 2</td>
<td>0.84232</td>
<td>26.1952</td>
</tr>
<tr>
<td>Model 3</td>
<td>0.74868</td>
<td>22.8974</td>
</tr>
<tr>
<td>Model 4</td>
<td>0.75919</td>
<td>22.0925</td>
</tr>
<tr>
<td>Model 5</td>
<td>0.83647</td>
<td>25.9400</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.85752</td>
<td>24.9883</td>
</tr>
</tbody>
</table>

4.3.1 Limitations

In order to gauge the performance of the proposed GAN model, we conducted a qualitative assessment on images that were dissimilar to any of those encompassed within the training set. Within this evaluation, certain image samples underwent a process of facial area reconstruction by the proposed GAN model, resulting in distorted outcomes. These distortions could potentially be attributed to factors such as the network’s failure to facilitate gradient propagation across all layers. The detailed presentation of these distorted images generated by the proposed GAN model, juxtaposed against their corresponding ground truths, can be observed in Fig. 10. It is pertinent to note that despite the distortions, the generated images remain identifiable by human observers, predominantly due to the preservation of significant facial features.

4.3.2 Face recognition challenge

To assess the capability of the proposed method in generating images that closely resemble actual facial images, we conducted experiments involving face recognition using edge devices compatible with deep learning-based classifiers, specifically MobileNet [26] and NasNet [27]. This evaluation was carried out with the utilization of paired images: genuine facial images and synthetic facial images generated by the proposed GAN model.

For the purpose of training and testing the classifiers, a dataset was meticulously curated, comprising 1000 real facial images and an equivalent count of plausible facial images. This dataset was divided into an 80:20 ratio for training and testing purposes. The performance assessment of the deep learning classifiers was carried out through metrics including Accuracy, Precision, Recall, and F-1 Score, as explicated by equations (6–9). The outcomes of these performance evaluations for face recognition, employing deep learning classifiers, are comprehensively detailed in Table 5.

\[
\text{Accuracy} = \frac{TP + FP}{TP + TN + FP + FN} \tag{6}
\]
Precision = \frac{TP}{TP+FP} \quad (7)

Recall = \frac{TP}{TP+FN} \quad (8)

F-1 \text{ Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)

In the context of the classification task aimed at distinguishing between authentic and credible faces, the MobileNet classifier showcased a precision of 96.25%, whereas the NasNet classifier achieved an accuracy of 94.50%. Moreover, impressive results were obtained across a spectrum of performance metrics. These findings emphasize that the synthetic faces produced by the suggested model exhibit discernible attributes, enabling them to be distinguished from genuine faces. As delineated in Table 5, the obtained results affirm that the images produced by the proposed UNET++ and CoGAN-based Pix2Pix model exhibit accuracy and possess the potential to be utilized in face recognition systems for the purpose of discerning between genuine and fabricated faces.

To evaluate the generalization of the proposed method, we tested the proposed method on test images of faces with masks extracted from the face mask detection (FMD) dataset [23] and evaluated for SSIM and PSNR metrics. The quantitative and qualitative results of the test are presented in Table 6 and Fig. 11.

### Table 5
Deep learning classifiers results for facial recognition

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F-1 Score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNet</td>
<td>Real</td>
<td>96%</td>
<td>96%</td>
<td>97%</td>
<td>96.25%</td>
</tr>
<tr>
<td></td>
<td>Plausible</td>
<td>94%</td>
<td>95%</td>
<td>95%</td>
<td></td>
</tr>
<tr>
<td>NasNet</td>
<td>Real</td>
<td>93%</td>
<td>94%</td>
<td>94%</td>
<td>94.50%</td>
</tr>
<tr>
<td></td>
<td>Plausible</td>
<td>94%</td>
<td>95%</td>
<td>95%</td>
<td></td>
</tr>
</tbody>
</table>

For the test conducted to determine the generalization of the proposed method, an SSIM value of 0.90872 was achieved that indicates for random masked faces a significant structural similarity is retained by the proposed method. Further, for the PSNR metric, it achieved a value of 29.4723 that indicate a significant amount of signal in images is retained over noise thus, making the proposed method capable of reconstructing facial area behind masks with significant accuracy. Since, the proposed method works on the principle of image-to-image translation therefore, better results can be obtained in paired images with

---

Table 6
Quantitative results on FMD dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>SSIM</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNET++ and CoGAN (Proposed)</td>
<td>0.90872</td>
<td>29.4723</td>
</tr>
</tbody>
</table>

---
the availability of ground truths and masked faces. Furthermore, to gauge the validity of the proposed method in removing face masks from low resolution images, we trained the proposed method with a few images of faces with synthetic masks captured from distance and low-resolution camera and tested for its performance. The results for the distant and low-resolution images in removing and reconstructing the face area behind masks is presented in Fig. 12.

4.4 Comparison with related work

To conduct a thorough comparative analysis between the proposed method and recent contributions in the field, we subjected benchmark face mask removal techniques to training and testing using the curated dataset. The prevailing trends in face mask removal approaches predominantly involve the utilization of GANs, albeit with variations in architectural designs. The comprehensive breakdown of this comparative analysis is provided in Table 7.

![Table 7](image)

Comparative analysis with previous studies

<table>
<thead>
<tr>
<th>Work</th>
<th>Method</th>
<th>SSIM</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ud Din et al. [11]</td>
<td>GAN employing a single generator and dual discriminators</td>
<td>0.85237</td>
<td>25.3752</td>
</tr>
<tr>
<td>Jiang et al. [13]</td>
<td>Single generator and discriminator-based GAN</td>
<td>0.89981</td>
<td>28.9085</td>
</tr>
<tr>
<td>Farahanipad et al. [14]</td>
<td>Cycle-GAN</td>
<td>0.86314</td>
<td>26.1518</td>
</tr>
<tr>
<td>Ours</td>
<td>UNET ++ and CoGAN-based Pix2Pix</td>
<td>0.91115</td>
<td>31.7251</td>
</tr>
</tbody>
</table>

The outcomes presented in Table 7 distinctly demonstrate that the proposed approach outperforms the GAN architectures outlined in the related studies, as evidenced by superior SSIM and PSNR metric scores. In particular, the method put forth demonstrates enhancements spanning from 1.2–5.9% for SSIM and 2.82–6.35% for PSNR metrics, when contrasted with the GAN architectures advocated in earlier studies. It’s worth highlighting that the technique advanced by Jiang et al. [21], which hinges on a single generator and discriminator, bears striking resemblance to the framework of the suggested approach. Nonetheless, the results continue to underscore the superiority of the proposed method in terms of structural similarity and signal preservation over noise, surpassing other GAN-based methodologies when evaluated on the dataset employed within this study.

5. Conclusions and Future Work

Generative Adversarial Networks (GANs) have emerged as a potent instrument for crafting realistic images, audio, and videos, with applications proliferating across various sectors like healthcare, design, and computer gaming. In this research, we present an innovative approach that amalgamates UNET ++ and CoGAN-based Pix2Pix GAN for the task of revealing concealed faces. The UNET ++ architecture is
meticulously crafted using a fusion of convolution layers, convolution transpose layers, max pooling layers, and batch normalization layers, each equipped with distinct kernel sizes and strides. Our devised method not only surpasses several state-of-the-art models but also exhibits impressive proficiency in removing masks. The incorporation of a modified UNET++ model as the generator notably enhances the overall performance of our GAN, particularly reflected in elevated SSIM and PSNR scores. A striking attribute of our model is its adaptability in eliminating various mask types, transcending the constraints of specific training masks.

Furthermore, we have curated an original dataset tailored for paired image-to-image translation, which uniquely encompasses images of individuals within the criminal domain, augmenting the dataset's relevance to real-world scenarios. In the future, our research direction could focus on harnessing the proposed GAN model for tasks like identity recognition and preservation. This might involve the utilization of models such as StyleGAN to infuse distinctive aesthetic characteristics into generated facial images, thereby broadening the horizon of potential applications.

Declarations

Availability of data and materials

1. The original Bollywood Celebrity Face dataset is available at: https://www.kaggle.com/datasets/havingfun/100-bollywood-celebrity-faces.

2. The created dataset and code of the UNET++ and CoGAN method is available on request.

Competing interests

The authors have no competing interests to disclose.

Funding

The authors express their gratitude to the All India Council of Technical Education for providing financial support for this study. This research has been sponsored under the Research Promotion Scheme of AICTE, India, with reference file number 8-108/FDC/RPS(POLICY-1/2019-20).

Authors’ contribution

Akhil Kumar (AK) contributed to methodology design, implementation, formal analysis and writing and editing original and final manuscript. Divyam Gupta (DG) contributed to methodology design, implementation and formal analysis. Manisha Kaushal (MK) contributed to editing and reviewing original and final draft of the manuscript. Akashdeep Sharma (AS) supervised the entire work with additional responsibility of conceptualization and methodology design.

Acknowledgements
References


Figures

Figure 1
Dataset description

Figure 2
Detection of face mask regions via the tiny YOLO v4-SPP detector

Figure 3

Dataset creation process
Figure 4
Architecture of proposed model

Figure 5
Discriminator model
Figure 6

UNET++ architecture
Figure 7

Plots for loss

a) Generator loss  b) Discriminator loss  c) Proposed method loss
Figure 8

Images generated by proposed GAN for faces without mask
Figure 9

Plots for SSIM and PSNR values with custom GANs

a) SSIM values  b) PSNR values
Figure 10

Failed examples with the proposed GAN model
Figure 11

Qualitative results of test on FMD dataset

Figure 12

a) Face with mask

b) Generated face without mask
Proposed method results on low-resolution images