HiMFR: A Hybrid Masked Face Recognition Through Face Inpainting

Md Imran Hosen[1], Md Baharul Islam[1,2*]

[1]imranjucse21@gmail.com, [2]bislam.eng@gmail.com,

[1]Department of Computer Engineering, Bahcesehir University, Istanbul, Turkey

[2]College of Data Science & Engineering, American University of Malta, Bormla, Malta

Abstract—To recognize the masked face, one of the possible solutions could be to restore the occluded part of the face first and then apply the face recognition method. Inspired by the recent image inpainting methods, we propose an end-to-end hybrid masked face recognition system, namely HiMFR, consisting of three significant parts: masked face detector, face inpainting, and face recognition. The masked face detector module applies a pretrained Vision Transformer (ViT_b32) to detect whether faces are covered with masked or not. The inpainting module uses a fine-tune image inpainting model based on a Generative Adversarial Network (GAN) to restore faces. Finally, the hybrid face recognition module based on ViT with an EfficientNetB3 backbone recognizes the faces. We have implemented and evaluated our proposed method on four different publicly available datasets: CelebA, SSDMNV2, MAFA, Pubfig83 with our locally collected small dataset, namely Face5. Comprehensive experimental results show the efficacy of the proposed HiMFR method with competitive performance.

Index Terms—Face detection, Face recognition, Vision Transformer, Face reconstruction, Inpainting, EfficientNetB3

I. INTRODUCTION

Face recognition has gained a lot of attraction and is still one of the most prominent research fields in computer vision and pattern recognition. It has become a trend for its practical and commercial applications, including face attendance system, face access control, mobile payment through face authenticity, and so on [1]. Many works have been reported while deep learning-based work [2]–[8] with the efficient loss function has shown promising recognition results. Weiyang Liu et al., [2] proposed SphereFace model: angular-based softmax loss function, which allows convolution neural network (CNN) to learn angularly discriminative features. However, many approximations are required to compute this loss that causes unstable network training. To solve training instability, Jiankang et al., [4] proposed ArcFace (Additive Angular Margin Loss) model to improve the discriminative ability of the model and stabilize the training process.

A significant face portion, including the nose and mouth, is covered with a mask that shortens the description parts of the face. Thus, the extension of existing face recognition methods for masked face recognition can significantly reduce the performance. According to a recent National Institute of Standards and Technology (NIST) [9], face recognition system performance drop nearly fifty percent when faces cover with a mask, and false negatives increase. In contrast, false positives were unchanged or slightly reduced [10]. Several unconventional approaches have recently been introduced for masked face recognition with different techniques such as discarded occlusion parts [5], [6], [8] and combining training (both mask and unmask) [7].

To recognize the masked face through discarding the mask part, the first mask portion is discarded and then combined with the applicable dictionary (set up by taking advantage of the distinctions between the top conv elements of hidden and visible face pairs) items. It is multiplied with the proper features to neglect the features that come from the masked part in recognition [5], [6], [8]. This technique's advantage is that it needs less time to train since it works with comparatively fewer data. However, this method provides poor results for mixing applications (masked and unmasked) due to partial visibility, which leads to a drop in information while dealing with unmasked faces. For example, while [8] got 95.07% accuracy for masked face, their performance reduced to 91.10% for mixture application. In [7] throughout the training system, both masked and unmasked datasets are shuffled one by one with the usage of the identical seed, and they decide the newly selected face image coming from the masked or unmasked data with 50% probability. However, this technique achieves good accuracy but requires a lot of training data with high computational costs.

Additionally, it is necessary to identify the masked face accurately and timely to recognize a masked face. Though some CNN based methods [11]–[13] have achieved competitive performance. There are some challenges to recognizing masked faces, including (i) the necessity for a highly accurate and efficient masked face detector, ii) partial visibility that reduces the description of the face, and iii) the necessity for a robust and computationally efficient face recognition model. To address the challenges mentioned earlier, we integrated three major modules: masked face detection, occlusion restoration or inpainting based on Generative Adversarial Network (GAN), and finally, face recognition using a hybrid Vision Transformer (Vision Transformer with EfficientNetB3 [14] backbone). The contributions of this paper are summarized as follows:

- A hybrid face recognition module based on the Vision Transformer integrated with a masked face detector and a face inpainting module has been proposed. Our HiMFR can automatically detect and remove the mask and restore the occlusion part to recognize the person.
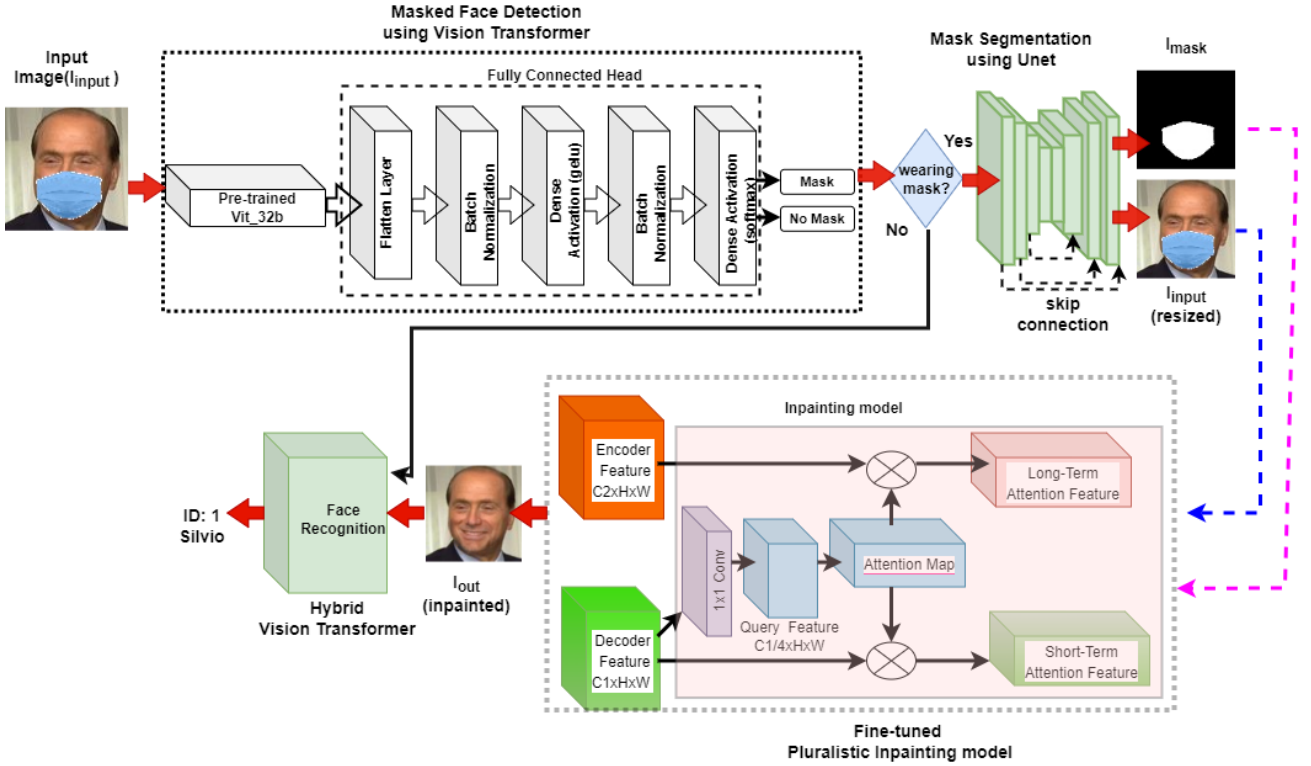
Fig. 1. The overall architecture of the proposed hybrid face recognition model that performs in three major modules: (i) masked face detector which determines face covered with mask or not, (ii) face restoring or inpainting module that is applied to restore the face, if the face is covered with mask and (iii) face recognition module.

- We propose a masked face detection module using the ViT_b32 with a fully connected (FC) head that provides competitive performance compared to different pre-trained models.
- To make the restoration faces convenient and realistic, we have fine-tuned the Pluralistic Image Completion (PIC) [15] method that can restore a large face occlusion region with more than one possible outcome.
- Our proposed HiMFR method is tested with four publicly available dataset including CelebA [16], SSDMNV2 [11], MAFA [17], Pubfig83 [18] and our collected Face5 dataset. The experimental results show promising performance to detect and recognized masked face compared to the state-of-the-art methods.

II. PROPOSED HiMFR METHOD

Our HiMFR model can be divided into masked face detection, face inpainting, and face recognition. Firstly, the masked face detection module determines whether the image contains a face mask or not. The intermediary face inpainting module is performed to restore the face before the face recognition module. The overall architecture of the proposed HiMFR model shows in Fig. 1.

A. Masked Face Detector

Our mask face detector module takes I_{input} as input, and provides a decision whether the mask is present or not in the face. The difference between a traditional face detector and our masked face detector is that the face detector determines the face portion. In contrast, our mask face detector provides a binary decision about mask detection. To reduce the computational cost and make the train faster, we have used transfer learning where pre-trained ViT_b32 has been used as a feature extractor. It trained on ImageNet21k [19] dataset which consists of around 14 million images. In the fine-tuning process, the last layer (fully connected head) of the ViT has been chopped, and a fully connected new head has been added to the pre-trained ViT_b32. The fully connected layer follows up on a packed input where all neurons are associated with each input. The FC head consists of a flattened, two batch normalization layers and two dense layers as shown in Fig. 1. We have used Categorical Cross-Entropy (CE) and Rectified Adam as loss functions and optimizers. The Categorical Cross-Entropy loss function shown in Eq. 1 is used in multi-class classification tasks where one predictable output belongs to one of many possible categories. Additionally, this function is intended to measure the distinction between two likelihood distributions.

$$CE = \sum_{n=1}^{n} y_i \log \hat{y}_i \qquad (1)$$

where n is the number of output size. $\hat{y}_i$ is the predictable output of ith target value of ith sample.
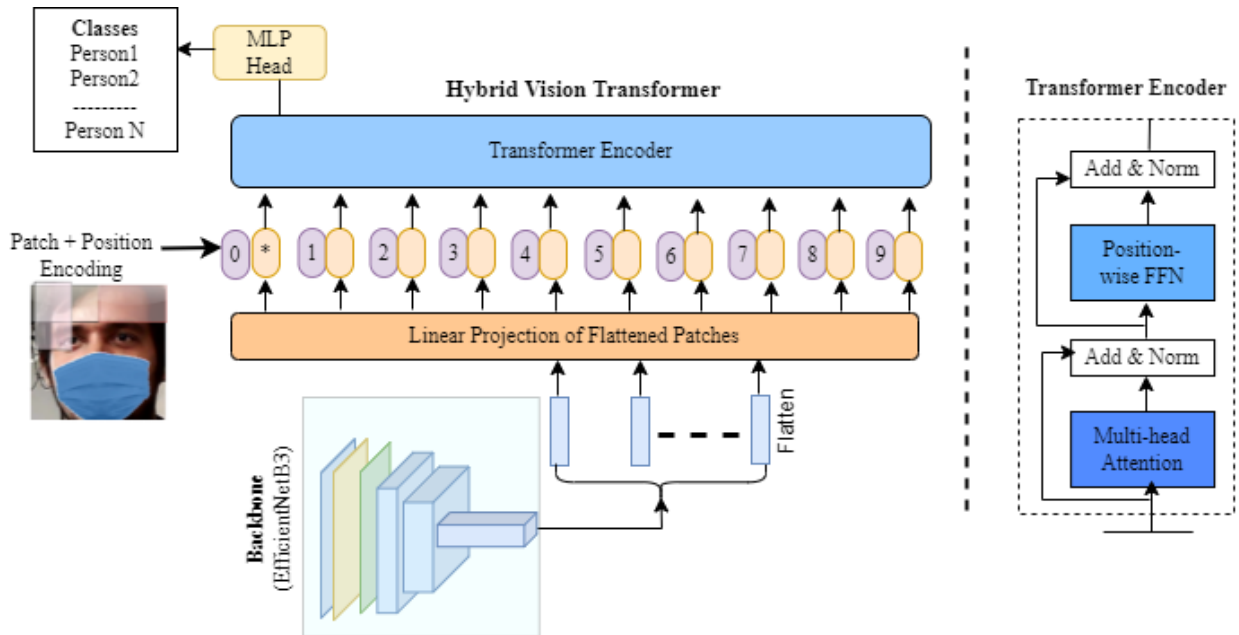
Fig. 2. The overall architecture of our proposed hybrid face recognition module is based on the ViT transformer and EfficientNetB3 as a backbone. Token has been generated with sliding patches rather than overlapping to increase performance. The encoder part is a standard Transformer encoder.

B. Face Inpainting

We have fine-tuned the pluralistic image completion (PIC) model initially proposed by Zheng et al. [15] which can produce more than one plausible outcome. The benefit of this model is that one can verify the person's identity in possible ways. Assume there is an original image (I_g) whose quality has been degraded due to some missing pixels. This degradation is represented by I_m (the masked partial image). I_c is also defined as the ground truth hidden pixels' complement partial image. The image completion method generates the missing pixels in a deterministic way. As a result, it is possible to reconstruct only one image. However, the authors used a non-deterministic way to generate multiple possible images in the PIC method [15]. Though the original PIC model can create various images for each person, we have developed a single image based on their discriminator scores in our approach. Additionally, one can verify the person with diverse possible outputs shown in Fig. 3. The three possible image completion results are shown by wearing the synthetic and real-world masks. The PIC model needs the binary mask to perform image generations. To segment the mask from the face, we used a pre-trained segmentation model [20].

C. Face Recognition

To recognize faces after occlusion removal, we have implemented a hybrid ViT that was initially proposed by Dosovitskiy et al. [21] with EfficientNet [14] that employs a revolutionary version scaling strategy. It uses a simple but powerful compound coefficient to scale up CNN's more structured manner. Traditional techniques use freely scaled network dimensions such as width, intensity, and resolution, while EfficientNet scales each variable uniformly using a
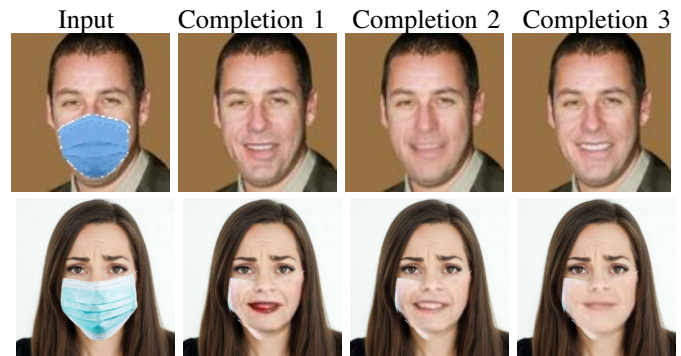

Fig. 3. The diverse and plausible outputs from the single input. First input from CelebA [16] and second input from real world MAFA [17] dataset.

predetermined set of scaling factors. We chopped the last layer of EfficientNet and added it (without the head) to the ViT. We have frozen the EfficientNet weights and implemented the sliding patches technique [22] to overlap the patches to increase the accuracy. Our model follows the ViT architecture shown in Fig. 2.

The overall architecture of the Vision Transformer model is as follows: (i) creating patches from an image (fixed sizes), (ii) making the image patches flat, (iii) creating lower-dimensional linear embeddings, (iv) adding positional embeddings, (v) feeding the sequence into the transformer encoder as an input, (vi) image labels use to train the ViT model, which is fully supervised shown in Eq.2 and Eq.3. The first step is the ViT separates an image into square patches in a grid and flattens the patch into a single vector. After that, learnable position embeddings enable the model to understand the image structure. In our implementation, hyperparameters such as learning rate, number of heads, transformer layer, etc., have been set. Then,

we augmented our data, and Multi-Layer Perceptron (MLP) was defined. The two-dimensional neighborhood structure is only utilized once in the model to divide the picture into patches at the model starting and to fine-tune the position embedding for images of varying resolutions [21].

$$z_0 = [Y_{class}; Y_p^1 E; Y_p^2 E; .....; Y_p^N E] + E_{pos}, \qquad (2)$$

$$z'_l = MSA(LN(z_{l-1})), l = 1...L$$
$$z_l = MLP(LN(z'_l)) + z'_l, l = 1...L \qquad (3)$$
$$y = LN(z_L^0)$$

where, Y_p represents flattened patches, $Y_p E$ indicates output patch embedding and $Y_{class} = z_0$ is a learnable embedding. MSA stands for multiheaded self-attention, and LN is Layernorm. z_0 serves as the image representation of y.

After defining MLP, we have implemented patch creation and patch encoding layer. By projecting a patch into a vector of size projection dimension, the Patch Encoder layer linearly converts it. It also adds to the projected vector a learnable position embedding. Our ViT model comprises 2 Transformer blocks, each of which uses the 8 Multi-Head Attention layer as a self-attention mechanism for the patch sequence. The Transformer blocks generate a tensor with a projection dimension, batch size, and several patches, then processed via a classifier head with a softmax activation function to generate the final probabilistic result.

III. DATASET AND EXPERIMENT

A. Dataset

To evaluate the efficiency of the HiFMR method and compare with state-of-the-art methods, we conduct experiment with four publicly available datasets, including CelebA [16], SSDMNV2 [11], MAFA [17], Pubfig83 [18] and our locally collected Face5 dataset. Data has been pre-possessed (e.g., cropping) to achieve better performance from each module. For instance, to train the image inpainting module, it needs to be higher quality with the proper size. Besides that, the image should contain only the face, so cropping is also necessary. The image in CelebA dataset [16] cropped into size 256×256 of the face. Then we create simulated mask [23].

CelebA [16] dataset contains 200K celebrity images which have large quantities, diversities, and rich annotations. Two other publicly available datasets, e.g., SSDMNV2 [11] contains around 11k data (5500 are mask and others are unmasked face), and MAFA [17] contains 6k data (3k with mask and 3k unmask face) have been used to compare the masked face detection module with state-of-the-art methods. We collect 5k data (size 224×224) from 5 individuals focusing on the face. We follow the data collection procedure by providing consent to the subjects. We have used a 2 MP camera resolution frame rate at 30 (frames per second) with the size of 480×640 pixels. We have also used Pubfig83 [18] (consist of 83 classes and each class contains around 100 images) dataset to train and testing of our face recognition module.

B. Experimental Setup

Experiments are conducted in windows 10, 32GB RAM, and one GPU (NVIDIA Geforce RTX 2070). To train and test each module, we split the dataset into 80% and 20% for training and testing. To train the masked face detector module, we have used processed CelebA unmask and synthetically created mask data (size 224×224). The term synthetically indicates that the data initially was unmasked faces, and then we put the same size mask artificially on the faces. We have run the experiment for five epochs and batch size 16 with Rectified Adam optimizer and a 0.0001 learning rate. We have run the image inpainting module for 150 epochs where Adam optimizer has been used with a learning rate of 0.0001 and 256×256 image size. We have used our Face5 dataset for face recognition module training and run it for 10 epochs with num_heads 4, batch size 2, transformer layers 2, num_heads 8, image size 224×224. Additionally, the Adam optimizer has been used with a fixed learning rate of 0.0003.

C. Evaluation Matrix

All modules of the proposed HiMFR method are separately evaluated. Accuracy, Precision, Recall, and F1 Score are calculated shown in Eq. 4 - 7.

$$Precision = \frac{TP}{(TP + FP)} \qquad (4)$$

$$Recall = \frac{TP}{(TP + FN)} \qquad (5)$$

$$Accuracy = \frac{TP + TN}{(TP + FP) + (TN + FN)} \qquad (6)$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (7)$$

where TP, TN, FP and FN indicate true positive, true negative, false positive, and false negative respectively.

Besides, we have used Structural Similarity Index (SSIM) and Peak Signal to Noise Ratio (PSNR) to compare image inpainting results. The SSIM measures the similarity between two images by comparing three measurements between images: luminance, contrast, and structure. Higher SSIM (up to 1) indicates the output image maintains its original structure. On the other hand, PSNR computes the correlation between the obtained output and the input. A higher PSNR value indicates better quality in the performance.

TABLE I
COMPARISON OUR MASKED FACE DETECTION MODULE WITH DIFFERENT PRE-TRAINED MODELS ON CELEBA [16] DATASET.

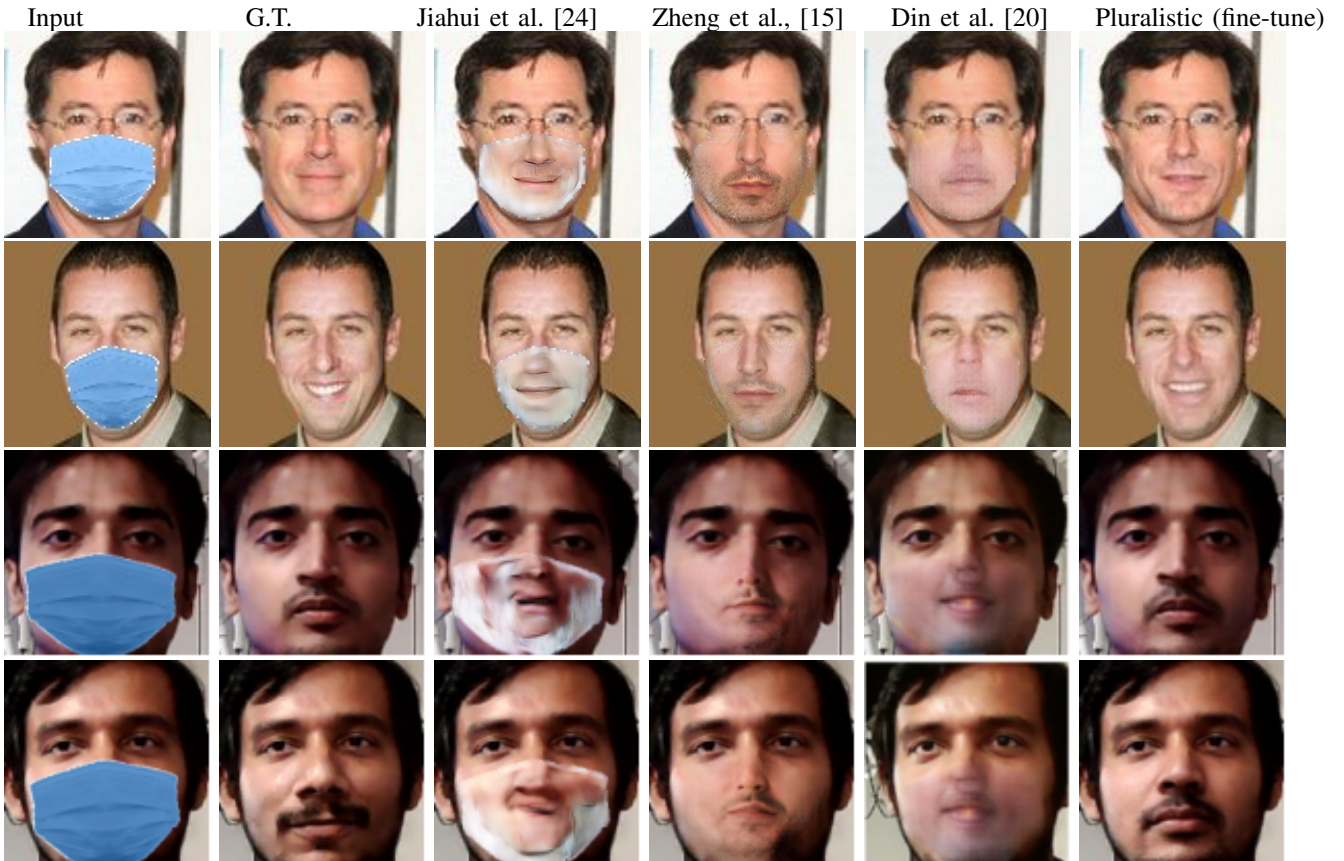| Architecture	Resnet50	MobileNetV2	VGG16	VGG19	HiMFR
Accuracy	99.03%	97.04%	96.56%	96.85%	99.90%

Fig. 4. Qualitative comparison of HiMFR fine-tune inpainting module with state of the art methods: on CelebA [16] (top 2 rows) and Face5 (bottom 2 rows) datasets.

TABLE II
COMPARISON THE PERFORMANCE OF THE HIMFR MASKED FACE
DETECTOR WITH STATE-OF-THE ART METHODS ON TWO DATASETS.

| Methods | Dataset | | Accuracy |
	MAFA [17]	SSDMNV2 [11]	
Sethi et al., [13]	✓	X	98.27%
Ahmed et al., [12]	✓	X	93.94%
Nagrath et al., [11]	X	✓	92.64%
Our HiMFR	✓	X	**99.04**%
Our HiMFR	X	✓	**99.32**%

TABLE III
QUANTITATIVE COMPARISON OF HIMFR INPAINTING MODULE WITH
STATE-OF-THE-ART METHODS IN TERMS OF PSNR AND SSIM METRIC.

Method	PSNR	SSIM
Yu et al., [24]	32.15	0.74
Zhen et al., [15]	33.34	0.79
Din et al., [20]	32.67	0.78
Our HiFMR	**35.92**	**0.90**

IV. EXPERIMENTAL RESULTS

A. Masked Face Detection Performance

The HiMFR masked face detector module is time-efficient and provides state-of-the-art performance with the minimum number of iterations shown in TABLE I. The Resnet50 achieved comparatively higher accuracy than the other pre-trained models (MobileNetV2, VGG16, and VGG19) in 5280 iterations, while our module achieved slightly higher accuracy within 1760 iterations. It indicates the faster performance of our HiMFR face detection module. To make a fair comparison, we have compared the HiMFR masked face detector module with the state-of-the-art methods demonstrated in TABLE II. Our module achieves higher accuracy compared to state-of-the-art techniques. Sethi et al., [13] acquired 98.27% accuracy near to our method.

B. Face Inpainting Performance

The quantitative evaluation of pluralistic image completion is challenging since it provides various possible solutions for a single masked image. However, to make the comparison, we took the single output from plausible generated outputs based on their discriminator score. We have conducted the comparison on the mixed dataset (CelebA [16] and Face5) in terms of Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) as shown in TABLE III. Fine-tune pluralistic inpainting achieved better performance than the original pluralistic [15] and other state-of-the-art methods [20], [24]. Additionally, we qualitatively compare the fine-tuned inpainting results with state-of-the-art methods, as presented in Fig. 4. Pluralistic [15] method generated outcome better than the [20], [24] but fine-tuned pluralistic image completion

generates a more realistic output with fine details and structure. Moreover, to test the applicability of our module in the real world, we have tested our module on the real-world masked MAFA dataset [17] shown in Fig. 3 (bottom row). As shown, our module produces realistic output. However, a small portion of masked is present. It was due to improper segmentation mask from the face.

C. Face Recognition Performance

To demonstrate the performance of our face recognition module, we have trained our face recognition module on Face5 dataset (five classes: each class contains 1000 image of each individual). The Receiver Operating Characteristic (ROC) curve is plotted using a testing set which consist of 200 images (100 masked faces and 100 unmasked), as shown in Fig. 5. Each color represents the different classes and the area under the curve (AUC) are 1.00, 0.99,0.99,0.96 and 0.98, indicating satisfactory performance. The performance matrix of our proposed masked face recognition system is demonstrated in TABLE IV for Face5 dataset (five classes and each class represents a individual person ($0 \sim 4$)). HiMFR achieved a competitive performance with 95% accuracy.

Additionally, to compare the performance of HiMFR with state-of-the-art methods, we adopted four [2]–[4], [8] deep learning-based face recognition methods. We tested on the Face5 dataset both unmask and mixed (mask and unmask) as shown in TABLE V. The performance of the state-of-the-art methods dropped significantly for the mixed data, while our masked face recognition shows the robust performance. Besides, our module achieved 94.7% accuracy on Pubfig83 dataset [18], which shows the competitive performance of our module compared to the Savchenko et al., [25] on the same dataset. To see the effectiveness of the backbone, we have trained and tested it without backbone and noticed the face recognize module performance drops significantly.

TABLE IV
PERFORMANCE MATRICES FOR PROPOSED HiMFR ON FACE5 DATASET.

Class label	Precision	Recall	F-score
0	1.00	1.00	1.00
1	0.90	0.94	0.92
2	0.95	0.95	0.95
3	0.95	0.90	0.92
4	0.95	0.95	0.95
Accuracy	**0.95**		
macro avg	0.95	0.95	0.94
weighted avg	0.95	0.95	0.95

V. CONCLUSION

This paper introduced a novel hybrid face recognition through the face reconstruction that works in three sub-modules: masked face detector, face inpainting, and face recognition. A Vision Transformer performed masked face detection. We used pre-trained ViT_b32 as a feature extractor with a fully connected head that achieved higher accuracy with the minimum iterations and made the masked face detector
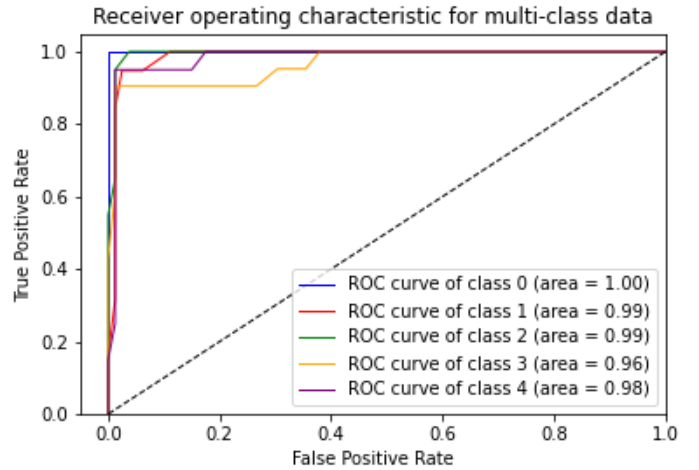

Fig. 5. Receiver operating characteristic (ROC) curve for presenting the relationship between the False and True Positive Rates.

TABLE V
COMPARISON OF OUR MASKED FACE RECOGNITION MODULE WITH STATE-OF-THE-ART METHODS ON FACE5 DATASET.

Model	Accuracy (Unmask)	Accuracy (masked and umasked)
Liu et al., [2]	97.00%	81.00 %
Geitgey et al., [3]	99.00%	83.00%
Deng et al., [4]	99.00%	90.00%
Junayed et al., [8]	-	91.10%
HiMFR	**99.00**%	**95.00**%

Fig. 6. An example of inaccurate face completion.

module time-efficient. The pluralistic image completion model was fine-tuned and used to reconstruct the hidden part of the face covered. Finally, we implemented a Vision Transformer with an EfficientNetB3 backbone-based face recognition module. Our method was evaluated quantitatively and qualitatively with different datasets. Although our HiMFR can recognize masked faces with high accuracy, it can be failed for inaccurate face completion by the image inpainting module (due to imprecise face mask tracking and segmentation). An example of the failure case shown in Fig.6 inaccurately. Our method is faster but does not perform in real-time. These limitations may consider in future works.

References

[1] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei *et al.*, "Masked face recognition dataset and application," *arXiv preprint arXiv:2003.09093*, 2020.

[2] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[3] A. Geitgey, "Face recognition documentation," *Release 1.2*, vol. 3, pp. 3–37, 2019.

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[5] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential siamese network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 773–782.

[6] W. Hariri, "Efficient masked face recognition method during the covid-19 pandemic," *arXiv preprint arXiv:2105.03026*, 2021.

[7] D. Montero, M. Nieto, P. Leskovsky, and N. Aginako, "Boosting masked face recognition with multi-task arcface," *arXiv preprint arXiv:2104.09874*, 2021.

[8] M. S. Junayed, A. Sadeghzadeh, and M. B. Islam, "Deep covariance feature and cnn-based end-to-end masked face recognition," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 1–8.

[9] M. Ngan, P. Grother, and K. Hanaoka, "Ongoing face recognition vendor test (frvt) part 6a: Face recognition accuracy with masks using pre-covid-19 algorithms," 2020-07-24 2020.

[10] S. Dharanesh and A. Rattani, "Post-covid-19 mask-aware face recognition system," in *2021 IEEE International Symposium on Technologies for Homeland Security (HST)*. IEEE, 2021, pp. 1–7.

[11] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria, and J. Hemanth, "Ssdmnv2: A real time dnn-based face mask detection system using single shot multibox detector and mobilenetv2," *Sustainable cities and society*, vol. 66, p. 102692, 2021.

[12] A. E.-A. Ahmed, N. Azim, M. Mahmood, and H. Alshammari, "A deep learning model for face mask detection," in *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 2', 10 2021, pp. 101–107.

[13] S. Sethi, M. Kathuria, and T. Kaushik, "Face mask detection using deep learning: An approach to reduce risk of coronavirus spread," *Journal of Biomedical Informatics*, vol. 120, p. 103848, 2021.

[14] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[15] C. Zheng, T.-J. Cham, and J. Cai, "Pluralistic image completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1438–1447.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[17] F. mask dataset (MAFA), "Kaggle.com, 2020. [online]," *Available: https://www.kaggle.com/andrewmvd/face-mask-detection.*, 2020.

[18] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 365–372.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[20] N. U. Din, K. Javed, S. Bae, and J. Yi, "A novel gan-based network for unmasking of masked face," *IEEE Access*, vol. 8, pp. 44 276–44 287, 2020.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[22] Y. Zhong and W. Deng, "Face transformer for recognition," *arXiv preprint arXiv:2103.14803*, 2021.

[23] A. Anwar and A. Raychowdhury, "Masked face recognition for secure authentication," *arXiv preprint arXiv:2008.11104*, 2020.

[24] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," *arXiv preprint arXiv:1801.07892*, 2018.

[25] A. Savchenko, "Efficient statistical face recognition using trigonometric series and cnn features," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3262–3267.