U-FA-MI: UNMASKING FOR FACIAL ATTRIBUTE PREDICTION ON MASKED IMAGES

Chaitra Jambigi, Gaurav Kumar Nayak, Anirudh Kannan, Anirban Chakraborty

Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India
{chaitraj, gauravnayak, anirban}@iisc.ac.in, anirudhkannan25@gmail.com

ABSTRACT

Face attribute prediction has applications in surveillance, face attribute manipulation, biometric recognition, etc. State-of-the-art deep models for facial attribute prediction assume the availability of full-face images. The COVID-19 pandemic has opened a new challenge for these models wherein the face gets partially covered due to a face mask. This work aims to find if existing face attribute prediction models are suitable for masked faces by creating masked versions of the existing face attribute dataset. We observe a drop in the performance of these models on masked faces, and to overcome that, we propose a novel Unmasking technique (dubbed as *U-FA-MI*) via a two-stage training pipeline. In the first stage, we propose a mask detection technique that detects the bounding box of the masked region on the prepared masked samples. In the second stage, we use an existing image inpainting technique, and we aid the training via our attribute consistency loss to generate the unmasked images. Our *U-FA-MI* framework can easily be plugged in with any attribute prediction network and does not require retraining. We experimentally validate our proposed method where we obtain a significant boost of 10.45% in average balanced accuracy on masked images of CelebA dataset.

Index Terms— Unmasking, Facial Attribute Prediction, Masked Images, Image Inpainting, COVID-19

1. INTRODUCTION

Face attribute prediction [1, 2, 3, 4, 5] aims to predict the common facial attributes such as gender, smiling face, the texture of hair, type of beard etc. Facial attribute analysis is an important task that has widespread use in applications such as Surveillance [6, 7], retrieval [8, 9], and Biometric recognition systems [10, 11, 12]. Along with security-based applications, Face attribute prediction is used in social media [13, 14] and in face manipulation and editing [15]. Due to its enormous applicability, there has been a growing interest in the research community to solve the problem of face attribute prediction.

Predicting face attributes is a challenging task due to complex face variations, different orientations, noise and partial occlusions. With the advent of Deep Convolutional Neural Networks and large scale face attribute datasets [16], lot of work has been done in the field of attribute prediction [1, 2, 3, 4, 5]. The state of the art deep face attribute prediction models extract feature vectors from the detected face and train classifiers for classifying the attributes. Zhong *et al.* [3] directly apply FaceNet and VGG-16 networks to capture attribute features of face image.

Although deep models provide significant boost in attribute prediction performance, new challenges have emerged since the global health crisis of COVID-19. Various preventive measures have been adopted worldwide to curb the spread of the disease. The use of facial masks covering the mouth and nose has proven to be the best measure to reduce the spread of the virus. This preventive measure poses a great impact on the reliance upon the existing attribute prediction systems, as masking the face occludes most of the face losing important visual cues. Various studies have shown the adverse effect of Covid-19 on face biometrics [17, 18, 19]. Barrero *et al.* [17] provided a detailed survey of the impact of COVID-19 on Biometrics. There also exist works [20, 21, 22] which study the impact of face masks for Face recognition task. However, currently none of the works study this issue from the perspective of attribute prediction task.

In this work, we first analyse the performance of face attribute prediction models on masked images, which are originally trained on images without masks. To verify this, we generate the masked faces from unmasked faces using MaskTheFace tool [23] which detects facial landmarks and puts a randomly choosen mask near mouth region. We observe a significant drop in the performance of these models on masked faces (refer Sec. 4.3). To overcome the performance degradation, we design a two stage training pipeline, which we call *U-FA-MI*. In the first stage, we propose a mask detector which detects the bounding box of the masked region using a Yolo [24] based detection mechanism. In second stage, we use an inpainting module [25] to unmask the masked regions. We empirically observe that simply plugging an inpainting module is not much beneficial as the inpainting module is trained to match the data distribution but may or may not preserve attribute information. To overcome this, we propose to use an attribute consistency loss that explicitly preserves the attribute information while generation (details in Sec. 3.4). We evaluate the performance of the attribute prediction model on the masked faces not only on average accuracy metric but also on average balanced accuracy which takes into account the data

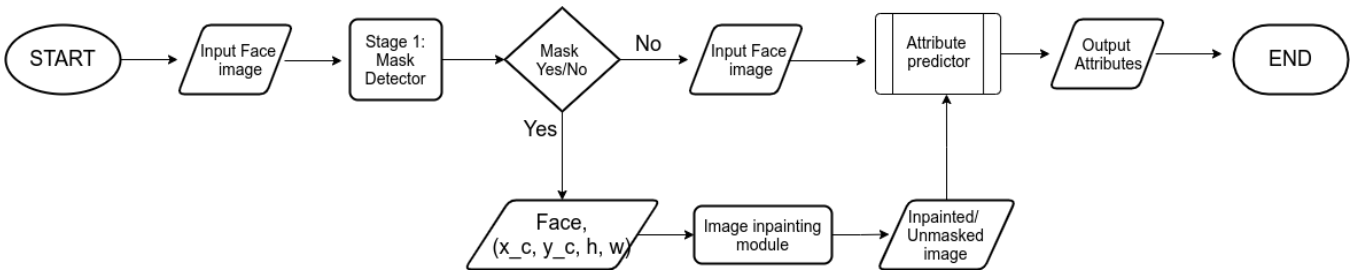

Fig. 1. Flowchart describing our overall pipeline for the *U-FA-MI* framework.

imbalance for rare attribute samples.

2. CONTRIBUTIONS

1. To the best of our knowledge, this is the first work that focuses on studying and improving the performance of face attribute prediction models on masked faces.
2. Our proposed method empowers existing face attribute prediction models that are not explicitly trained on masked faces to enhance their performance by allowing easy integration through prepending our module before the trained models.
3. The knowledge stored in attribute prediction models trained on images without masks is utilized through the proposed attribute consistency loss that helps to improve the image inpainting process resulting in better unmasked images and improved performance.
4. We demonstrate our method's efficacy qualitatively (Fig. 3), and empirically validate the models trained on images without mask. Our module is added on top of it that yields 10.45% and 3.4% improvement in avg. balanced accuracy over baseline on CelebA and LFW-A datasets, respectively.

3. PROPOSED APPROACH

We describe our proposed approach in the following sections. Sec.3.1 gives a detailed explanation on the flow of our pipeline, Sec.3.2 describes about the creation of masked images, Sec.3.3 describes about the first stage: mask bounding box detection, Sec.3.4 describes the second stage: image inpainting module, along with our attribute consistency loss.

3.1. *U-FA-MI* Framework

The main goal of our method is to empower the existing attribute prediction models to perform well on masked as well as unmasked images without retraining them. We describe the flow of our operations in the Flowchart in Fig1. We first pass the facial image through a mask detector, which detects the presence or absence of a face mask. If the mask is absent then we pass the image directly through the attribute prediction model. However, if there is a masked face, we pass the image through our pipeline which unmasks the masked region. Once we obtain the unmasked face image, we then give it to the attribute prediction model. Thus, our training pipeline helps to adapt the Face attribute prediction model on both masked and unmasked images without retraining the model.

3.2. Creation of Masked Images

We have large scale face attribute datasets: CelebA, LFW-A [16] for unmasked images, but we do not have their corresponding masked versions. Also, there isnn't any large scale masked face dataset with attribute annotations. Thus, we manually generate the masked version of the CelebA dataset, to train our pipeline. We use off the shelf tool 'MaskTheFace' [23] to generate the masked versions. The tool uses a dlib based face landmarks detector to identify the face tilt and six key features of the face necessary for applying mask. 'MaskTheFace' identifies the face within an image, and applies the user selected masks to them taking into account various limitations such as face angle, mask fit, lighting conditions etc. Following [23] we apply random masks of 3 types:cloth mask, surgical mask, N95 mask to create different variations of masked faces. For each image we randomly select a mask and generate a single new masked image. Fig.2 shows proposed framework which is discussed in subsequent sections.

3.3. Stage 1: Mask Detection

As a first step in our pipeline we detect if a given facial image contains a mask or not. We rely on an external Kaggle dataset [26] which contains masked and unmasked images with bounding box annotations. An i^{th} training sample has label $y_i = \{m_i, bbox_i\}$ having $m_i \in \{0, 1\}$, with 0 denoting absence and 1 denoting presence of mask. The real valued bounding box co-ordinates $bbox_i = (x_{ci}, y_{ci}, h_i, w_i)$ has center coordinates (x_{ci}, y_{ci}) with w_i and h_i denoting the width and height of the masked region respectively. For ease of reading, we avoid putting the subscript i for each sample. Let $D(.)$ denote the detection model. For a sample x, we have:

$$y, (x_c, y_c, h, w) = D(x) \tag{1}$$

We use Yolo-V3 [24] to detect and localise the mask region. However, bounding boxes in the dataset includes other features of the facial region as well (such as the eyes and nose),

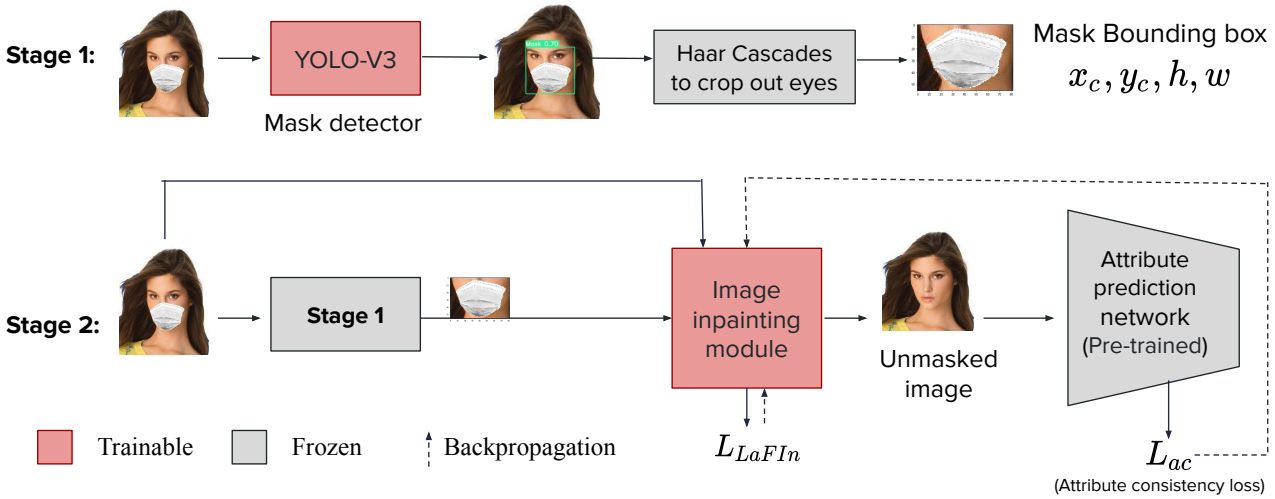

Fig. 2. Architecture diagram for the proposed *U-FA-MI* framework. Top row denotes the Stage 1 (Detection module) and bottom row denotes the Stage 2 (Image inpainting module) along with our Attribute consistency loss L_{ac}.

Haar cascades were used after the Yolo-V3 model to reduce the size of detected bounding box. The obtained bounding box (x_c, y_c, h, w) then accurately represents the face mask region. Once detector is trained in Stage 1, we freeze it's weights and use it as a pretrained detector in Stage 2.

3.4. Stage 2: Image Inpainting using Attribute consistency loss

Let $X \in \{x_1, x_2, \cdots, x_N\}$ denote the training samples of face attribute datasets which are masked as described in Sec.3.2. Let $I(.)$ denote the Image inpainting module. It takes input masked training sample x along with the masked region (x_c, y_c, h, w) which needs to be inpainted. We obtain an unmasked output image $\hat{x}$ after inpainting.

$$\hat{x} = I(D(x)) \tag{2}$$

Let $A(.)$ denote the attribute prediction model through which we pass $\hat{x}$ to get the prediction scores on unmasked images. We experimentally show in Sec.4.3 that the attribute prediction model performs better after unmasking of face images. However, it is worth noting that the Image inpainting model is trained to generate realistic face images with same distribution as input data. This does not guarantee preservation of facial attributes after generation. Thus, simply using a pretrained Image inpainting model may not be the best solution which we show in Sec.4.3. To overcome the above issue, we use the pretrained attribute prediction model $A(.)$ which has been previously trained on unmasked images, to transfer the attribute knowledge during generation training. Specifically, we add a new loss called as Attribute consistency loss L_{ac} which is the loss due to incorrect attribute predictions on unmasked images generated by inpainting model. Let M be the number of binary attributes. Let $\{y_1, y_2, ..., y_M\}$ denote the attribute labels where $y_i \in \{0, 1\}$. Since the task is of binary, multi-label classification, we use a binary cross-entropy loss function for each attribute during training following [2]. Let $\hat{y}$ denote the predicted attribute labels,

$$\hat{y} = A(\hat{x}) \tag{3}$$

Then L_{ac} loss can be defined as:

$$L_{ac} = -\sum_{k=1}^{M} [y_k \log \hat{y}_k + (1 - y_k) \log(1 - \hat{y}_k)] \tag{4}$$

$$L = \lambda_1 L_{ac} + \lambda_2 L_{LaFIn} \tag{5}$$

L_{LaFIn} is the Inpainting module loss as followed in [25]. Through attribute consistency loss, we pass the gradients caused by incorrect prediction of attributes to guide the inpainting process so as to reduce the L_{ac} loss. The main contribution of Attribute consistency loss is to guide the image generation to learn to preserve the attribute information as much as possible. We show in Sec.4.3 that we get a further boost in performance by adding attribute consistency loss with Image inpainting training.

4. EXPERIMENTS

4.1. Implementation Details

We perform experiments on two large scale face attribute datasets: CelebA, LFW-A [16]. We use the light-weight network SlimCNN [2] as our Attribute prediction model. The SlimCNN model is trained on the celebA dataset [16] with a learning rate of 0.001 and batch size of 128 for 15 epochs. The same hyperparameters are used to train the model on LFW-A [16]) for 35 epochs. We train Yolo-V3 [24] for 100

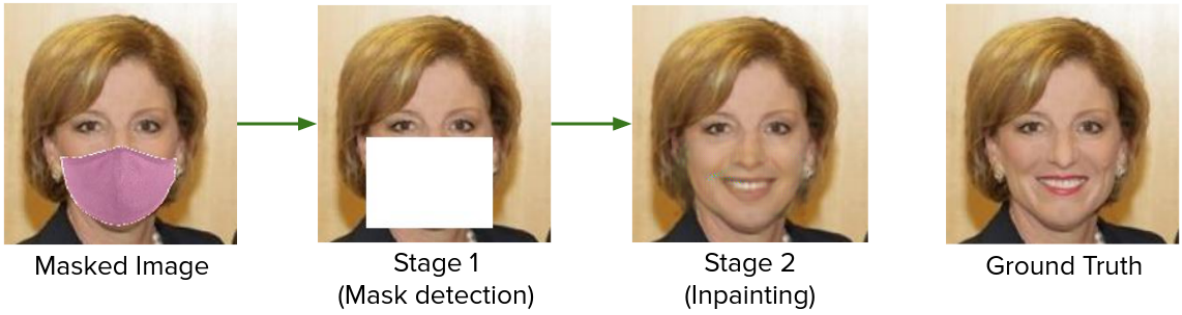

Fig. 3. Qualitative demonstration of our proposed approach. Once the masked images are prepared, they are sent to stage 1 (detection of masked region)and stage 2 (image inpainting trained via our proposed attribute consistency loss) sequentially.

epochs on a batch size of 16 following [26] on a Kaggle dataset which has masked and unmasked images found by scrapping Google images. The dataset details can be found at [26]. We use the state of the art image inpainting model, LaFIn [25] to inpaint the masked regions of the face. LafIn was trained for 776k iterations, after which our L_{ac} loss was added and the GAN was further trained till 1166K iterations. We give a weightage of 0.1 and 0.9 for λ_1 and λ_2 respectively.

4.2. Evaluation Metrics

As Facial attribute prediction is a binary classification task, it's accuracy on each attribute is defined as [27]:

$$Accuracy = \frac{(t_p + t_n)}{(N_p + N_n)} \tag{6}$$

where Np and Nn denote the number of positive and negative samples while and tp and tn denote the number of true positives and true negatives for the corresponding attribute. When dealing with class-imbalance data, the above accuracy is not a fair metric due to the bias of the majority class. Hence, balanced accuracy is defined as [27]:

$$Balanced\ accuracy = \frac{1}{2}(\frac{t_p}{N_p} + \frac{t_n}{N_n}) \tag{7}$$

We report the average performance of the above two metrics over all the attributes.

4.3. Results

As shown in Table 1, the SlimCNN model has an avg. accuracy of 89.84% and avg. balanced accuracy of 76.38% when tested on images without any facemask (celebA). When images with facemasks are tested on this model, the avg. accuracy drops to 82.14% and avg. balanced accuracy to 59.94%, owing to the fact that a facemask occludes multiple facial attributes. These results act as the baseline, and we adopt our training pipeline to improve the performance. Using the proposed *U-FA-MI* framework, avg. accuracy has improved by about 5.18% and balanced accuracy by about 7.58%. Furthermore, using our L_{ac} loss with *U-FA-MI* framework gives an additional improvement of about 6.75% on avg. accuracy and by 10.45% on avg. balanced accuracy compared to baseline. Similar trends can be seen in the performance of our proposed method on the LFW-A dataset in Table 2. We show qualitative results of our method in Fig.3.

Method	**Avg. Acc.**	**Avg. Balanced Acc.**
Images without mask (upper bound)	89.84%	76.38%
Masked faces (Baseline)	82.14%	59.94%
U-FA-MI (**Ours**)	87.32% (↑5.18)	67.52% (↑7.58)
U-FA-MI using L_{ac} loss (**Ours**)	88.89% (↑6.75)	70.39% (↑10.45)

Table 1. Performance of our proposed method when used on existing attribute prediction model (SlimCNN) on CelebA

Method	**Avg. Acc.**	**Avg. Balanced Acc.**
Images without mask (upper bound)	81.70%	72.99%
Masked faces (Baseline)	73.51%	63.77%
U-FA-MI (**Ours**)	76.62% (↑3.11)	66.91% (↑3.14)
U-FA-MI using L_{ac} loss (**Ours**)	77.15% (↑3.64)	67.17% (↑3.4)

Table 2. Performance of our proposed method as tested on existing attribute prediction model (SlimCNN) on LFW-A

5. CONCLUSION

We studied the generalisation ability of Face attribute prediction model on masked faces and observed a significant drop in performance. To alleviate this, we proposed an unmasking based framework (*U-FA-MI*). Our proposed attribute consistency loss furhter leads to gain in performance as validated experimentally over different datasets. As a future work, extending our framework to detect and inpaint other occluded facial regions for attribute prediction task would be an interesting direction to explore. The utility of our framework for other face related tasks can be another direction of research.

6. REFERENCES

[1] Fang Wang, Hu Han, Shiguang Shan, and Xilin Chen, "Deep multi-task learning for joint prediction of heterogeneous face attributes," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 173–179.

[2] Ankit Kumar Sharma and Hassan Foroosh, "Slim-cnn: A light-weight cnn for face attribute prediction," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 329–335.

[3] Yang Zhong, Josephine Sullivan, and Haibo Li, "Face attribute prediction using off-the-shelf cnn features," in *2016 International Conference on Biometrics (ICB)*. IEEE, 2016, pp. 1–7.

[4] Jianshu Li, Fang Zhao, Jiashi Feng, Sujoy Roy, Shuicheng Yan, and Terence Sim, "Landmark free face attribute prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4651–4662, 2018.

[5] Mahdi M Kalayeh, Boqing Gong, and Mubarak Shah, "Improving facial attribute prediction using semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6942–6950.

[6] Daniel A Vaquero, Rogerio S Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk, "Attribute-based people search in surveillance environments," in *2009 workshop on applications of computer vision (WACV)*. IEEE, 2009, pp. 1–8.

[7] Jongpil Kim and Vladimir Pavlovic, "Attribute rating for classification of visual objects," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 1611–1614.

[8] Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar, "Describable visual attributes for face verification and image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1962–1977, 2011.

[9] Siyu Xia, Ming Shao, and Yun Fu, "Toward kinship verification using visual attributes," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 549–552.

[10] Philipp Terhörst, Daniel Fährmann, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper, "Beyond identity: What information is stored in biometric face templates?," in *2020 ieee international joint conference on biometrics (ijcb)*. IEEE, 2020, pp. 1–10.

[11] Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei, and Stan Z Li, "Multi-label cnn based pedestrian attribute learning for soft biometrics," in *2015 international conference on biometrics (ICB)*. IEEE, 2015, pp. 535–540.

[12] Andrea F Abate, Paola Barra, Silvio Barra, Cristiano Molinari, Michele Nappi, and Fabio Narducci, "Clustering facial attributes: Narrowing the path from soft to hard biometrics," *IEEE Access*, vol. 8, pp. 9037–9045, 2019.

[13] Guo-Jun Qi, Xian-Sheng Hua, and Hong-Jiang Zhang, "Learning semantic distance from community-tagged media collection," in *Proceedings of the 17th ACM international conference on Multimedia*, 2009, pp. 243–252.

[14] Guo-Jun Qi, Charu Aggarwal, Qi Tian, Heng Ji, and Thomas Huang, "Exploring context and content links in social media: A latent space method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 850–862, 2011.

[15] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu, "Talk-to-edit: Fine-grained facial editing via dialog," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13799–13808.

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

[17] Marta Gomez-Barrero, Pawel Drozdowski, Christian Rathgeb, Jose Patino, Massimmiliano Todisco, Andras Nautsch, Naser Damer, Jannis Priesnitz, Nicholas Evans, and Christoph Busch, "Biometrics in the era of covid-19: challenges and opportunities," *arXiv preprint arXiv:2102.09258*, 2021.

[18] Marta Calbi, Nunzio Langiulli, Francesca Ferroni, Martina Montalti, Anna Kolesnikov, Vittorio Gallese, and Maria Alessandra Umiltà, "The consequences of covid-19 on social interactions: an online study on face covering," *Scientific reports*, vol. 11, no. 1, pp. 1–10, 2021.

[19] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper, "Unmasking face embeddings by self-restrained triplet loss for accurate masked face recognition," *arXiv preprint arXiv:2103.01716*, 2021.

[20] David Montero, Marcos Nieto, Peter Leskovsky, and Naiara Aginako, "Boosting masked face recognition with multi-task arcface," *arXiv preprint arXiv:2104.09874*, 2021.

[21] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou, "Masked face recognition challenge: The insightface track report," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1437–1444.

[22] Hanjie Qian, Panpan Zhang, Sijie Ji, Shuxin Cao, and Yuecong Xu, "Improving representation consistency with pairwise loss for masked face recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1462–1467.

[23] Aqeel Anwar and Arijit Raychowdhury, "Masked face recognition for secure authentication," 2020.

[24] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[25] Yang Yang, Xiaojie Guo, Jiayi Ma, Lin Ma, and Haibin Ling, "Lafin: Generative landmark guided face inpainting," *arXiv preprint arXiv:1911.11394*, 2019.

[26] Alexandra Lorenzo, "Maskdetection at yolo format," `https://www.kaggle.com/alexandralorenzo/maskdetection`, May 15, 2020.

[27] Ethan M Rudd, Manuel Günther, and Terrance E Boult, "Moon: A mixed objective optimization network for the recognition of facial attributes," in *European Conference on Computer Vision*. Springer, 2016, pp. 19–35.