maskedFaceNet: A Progressive Semi-Supervised Masked Face Detector

Shitala Prasad, Yiqun Li, Dongyun Lin, Dong Sheng
Visual Intelligence, Institute for Infocomm Research (I²R), A*STAR, Singapore
{shitalap, yqlin, lin_dongyun, dong_sheng}@i2r.a-star.edu.sg

Abstract

To reduce the risk of infecting or being infected by the recent COVID-19 virus, wearing mask is enforced or recommended by many countries. AI based system for automatically detecting whether individuals are wearing face mask becomes an urgent requirement in high risk facilities and crowded public places. Due to lacking of existing masked face datasets and the urgent low-cost application requirement, we propose a progressive semi-supervised learning method – called maskedFaceNet to minimize the efforts on data annotation and letting deep models to learn by using less annotated training data. With this method, the detection accuracy is further improved progressively while adapting to various application scenarios. Experimental results show that our maskedFaceNet is more efficient and accurate compared to other methods. Furthermore, we also contribute two masked face datasets for benchmarking and for the benefit of future research.

1. Introduction

Object detection (OD), being a fundamental problem in computer vision (CV), is a contemporaneous process of estimating the types and locations of object instances in an image or video frames. Similar to human vision system, CV has many real applications such as scene and document analysis [34, 9], riderless [1, 45], health care [41] and robotics [47, 39]. Since past few decades when OD was casted as a machine learning problem, several handcrafted features and classifiers were proposed [13]. But after the success of AlexNet [19], convolutional neural network (CNN) increased exponentially in CV applications.

One of the most focused CV problem is face detection (FD) which is considered as the basic step for any face associated applications such as face tracking and recognition [29, 12]. Since pentad, deep learning (DL) emerged to be a promising footstep in the field of face detection [51, 55]. However, the need of real-time FD with high accuracy in complex scenario is still a challenging due to occlusion, illumination, pose and scale variations.

As an exceptional case of OD, FD utilizes similar features and adopts many state-of-the-art OD methods as their backbone. These powerful CNN-based FDs attempt to address the above challenges to some extents by exploiting the feature maps [43] and computing extra information or by applying dense anchors [57]. But in spite of their success towards human-level for most of the images, an evident research gap still exist with those faces which are blur or occluded by 50% or more like the masked faces (Figure 1) or dark faces with dark masks (Figure 1f). In the recent COVID-19 pandemic, wearing mask is recommended to minimize the spread of the virus. In certain countries, it is compulsory to wear face mask when people visit to crowded public places. Wearing face mask is also common in places like health care centers and pharmaceutical labs. In above situations, the existing FDs mostly fails or performs poor because they are trained using easy face datasets where faces are almost frontal/profile and complete (or has less occlusion). As stated by Chi et al., today most of the FDs are focused to detect faces with high recall rate while ignoring their precision [7].

We believe training FDs with such easy images (Figure 1a) is not much useful in real-life scenarios, as mentioned above. On contrary, to collect huge real wild data with occlusions and then annotate them is a tedious and time consuming job. Sometimes, data annotation needs domain experts such as in biomedical tasks where it is really expensive to hire experts. Therefore, this paper focus on the real use cases where huge object-level annotation with high precision is an important challenge and progressively learning the object variations that keeps changing periodically/non-periodically. The challenges involved in masked face detection is shown in Figure 1 where different possible cases including faces with and without mask of different shape and color are highlighted in image.

Today’s digital world is full of unlabeled data that can help the networks in their learning process. Hence, the semi-supervised learning (SSL) algorithm utilizes the benefit from both labeled and unlabeled data. In classical SSL image classification, the model is updated using labeled and label-estimated from the model itself [32]. Therefore, the
key concept of SSL is to efficiently improve the model loss. In this paper, we investigated the effect of data size and recursive use of weight initialization in object detection SSL to improve the deep learning based FD. We propose a novel three-phase semi-supervised training strategy to efficiently detect faces with and without mask in public places. Inspired by [48], the proposed masked FD utilizes the unlabeled data for better weight initialization to ameliorate the performance index compared to the standard baseline approach (details are in Section 4). To summarize the contributions of this paper:

- We explore use of less data for training deep models with varying parameters to study the potential direction of progressive semi-supervised masked face detection. For this study, we introduce two real scenario datasets which include faces with and without mask: MASK-face-v1 and MASK-face-v2.

- We proposed a real-time light weighted deep detector, called maskedFaceNet which utilizes unlabeled dataset to boost the network performance.

- We conducted a comprehensive experimental analysis to verify the challenges involved in masked face detection using less annotated training data. We also examine the role of pseudo-labels in object detection task.

The main advantage of our proposed method is that it shares the knowledge to identify similar objects which reduces the effort of object-level annotation. The paper is organized in five sections. Section 2 describes the related works while Section 3 introduce the details of the proposed approach followed by extensive analysis and ablation studies on maskedFaceNet in Section 4. Finally, Section 5 concludes the paper and discusses the potential future directions of masked face recognition.

2. Related Work

As the problem definition stated in previous section, the masked face detection, similar to generic object detection [31], refers to localization of human faces with and without mask in an image. In this section, we cover various CNN-based OD algorithms followed by face detection methods and finally address a few state-of-the-art approaches in detecting masked faces.

2.1. Object Detection

CNN-based learning has deep impact in CV applications [31] such as education and surveillance [30] and others [31]. The traditional ODs were based on some hand-crafted features such as Haar features [24], scale-invariant feature transform [56] and feature pyramid [13]. These features needs to be engineered very carefully and they are very much application dependent. Recently, deep learning based ODs which adopt strong supervision in learning becomes dominant due to their excellent performance. ODs normally follow two major approaches: bottom-up and top-down, among which later is more common in deep models. Top-down approaches are further categorized into two: two-stage (Fast and faster R-CNN) and one-stage (YOLO and SSD) methods. Two-stage methods [37, 10] mainly focus on reducing the negative examples produced from the dense sliding windows, called anchors, while one-stage methods [28, 50] directly aims to predict results from anchors after feature extraction from the input image. Unlike two-stage approach, SSD framework gets benefit due to it’s higher inference efficiency and therefore attracts attention for real-time face detectors. In DL, all the variants of Faster R-CNN (Region-based Convolutional Neural Network), R-FCN (Region-based Fully Convolutional Network), SSD (Single Shot Detector) and YOLO (You Only Looks Once) are heavily dependent on huge training data which are manually annotated with objects and their local-
izations (e.g. ImageNet [11] and COCO [26]), which is very tedious job.

2.2. Face Detection

Under OD, face detection is one of the most important and challenging task which is grouped into three possible categories. The first category is boost-based FD which adopts boosted cascade Haar features [44]. SURF cascade [22] and Normalized Pixels Difference [23]. The second category is Deformable Part Model (DPM) based where deformation of faces are modeled. For example, Chen et al. proposed a joint detection and alignment FD in a single framework [3] while Ghiasi and Fowlkes proposed a joint detector that can handle face detection as well as key feature localization [14]. The third method is CNN-based which directly learn features from the input image, for example, CascadeCNN [21], Contextual Multi-Scale Region-based CNN (CMS-RCNN) [58], Supervised Transformer Network (STN) [3], MTCNN-based [52], Hyperface [35], YOLO-face [5] and Face-SSD [17]. CascadeCNN is a boosted exemplar-based FD while CMS-RCNN is unconstrained contextual multi-scale FD. In [33], Opitz et al. introduced a novel grid loss to deal with partial occlusion in face detection task. In contrast to this, Chen et al. proposed STN with a cascade CNN to address the challenge of huge face pose variation in real-world for face detection. Next, Ranjan et al. proposed CNN-based face detection and gender recognition [35]. In [17], Fully Convolutional Neural Network (FCNN) was used to detect multiple faces in a single image of different sizes. There are several other FDs proposed by various researchers that utilize advantages of two-stage and one-stage approaches such as FANet [51] and S3FD [54] but mostly perform poorly for masked faces or faces with more than 50% occlusion.

2.3. Masked Face Detection

In past literature, there is no much work related to masked face detection and therefore very limited articles are found. The first reported masked FD in wild image was by Ge et al. where they used locally linear embedding with CNN [14]. They also introduce a MAked Faces (MAFA) dataset with 30,811 Internet images. Each image in MAFA contains at least one face occluded by various types of masks which also includes faces covered with hand or other objects. The second recently published RetinaMask is more close to masked face detection objective where a subset of MAFA and WIDER [49] datasets with and without masked faces are created and tested to achieve an average precision rate of 92.65%. Here, the occlusion is not only mask but other objects too. Qiting Ye [50] proposed a novel framework using MTCNN [52] and VGG-16 [40] for masked face detection. With similar motivation, Lin et al. proposed a modified version of LeNet (MLeNet) for surveillance video masked face detection [25]. Chen et al. [6] used adversarial occlusion-awareness for face detection on MAFA dataset.

Other than these, in current COVID-19 pandemic many countries like France and companies such as SenseTime implemented their own system to monitor people wearing masks in public places. However, the research in this direction is still very limited and the methods or the trained models may not be able to work well or difficult to adapt to various application scenarios. Even researchers from NIST [4] found that the existing face recognition models fails as much as 50% of the time. Motivated with SSD object detector [28] and teacher-student model [48], we proposed a light weighted real-time maskedFaceNet to detect faces in wild scenario which requires less annotated training data and utilizes semi-supervised data for recalibration of weights. Our proposed model, similar to biological phenomenon of human vision system, uses a receptive field (RF) to increase the eccentricity of feature maps.

3. Proposed Methodology

In this section, we introduce our proposed single shot scale-invariant maskedFaceNet followed by the training strategy, the loss function and the implementation details.

3.1. Problem Definition

Generally in image classification task, with a given labeled image dataset, say \( D = \{(x, y)\} \), and an unlabeled image dataset, say \( U = \{(u)\} \), SSL aims to solve the following problem:

\[
\min_{\theta} \sum_{(x,y) \in D} L_{SL}(x, y, \theta) + \beta \sum_{(u) \in U} L_{UL}(u, \theta) \tag{1}
\]

where \( L_{SL} \) and \( L_{UL} \) represents the supervised and unlabeled loss, respectively. The value \( \theta \) is the total trainable parameters of the given model and \( \beta \) is the weight balancing parameter which is \( R_{>0} \). In notation, \( y \) represents the hard-label for image data \( x \) but their is no label for data \( u \), as it belongs to the unlabeled set, but \( D < U \). There are different ways to compute pseudo-labels for \( u \) and calculate the per-example unsupervised or semi-supervised loss \( L_{UL} \) proposed by various CV researchers [20, 46, 42]. In [38], Ren, Yeh and Schwing stated that not all unlabeled data are equal and therefore introduced per-example weights to compute \( L_{UL} \). This improves the performance of SSL algorithm but leads to the computational expense and there-
fore, they used influence function to deal with it. In this paper, following the SSL for image classification, we progressively used the unlabeled data for object detection. Unlike [38], we used varying confidence $\alpha$ for all $u \in U$. Specifically, instead of obtaining only the pseudo-label $\tilde{y}$ for $u$, we also consider the network’s confidence for the predicted label for it’s weight contribution, i.e., $\tilde{(y, \alpha)} = p_\theta(c|u)$. Thus, the above problem is redefined as:

$$\min_{\Theta} \sum_{(X,Y) \in D} L_{SL}(X,Y,\Theta) + \beta \sum_{(U,\hat{Y}) \in U} L_{UL}(U,\hat{Y},\Theta)$$

\hspace{1cm} (2)

where $(X,Y)$ and $(U,\hat{Y})$ are the input and output pairs respectively for labeled and unlabeled data where each set has five tuples, i.e., four coordinates with width and height and the face class $c$. In case of $\hat{Y}$ there is an extra score tuple $\alpha$ for fair $\Theta$ parameters learning. Thus, the weight importance $w_u$ for the unlabeled data is initialized as $\lambda \times \alpha$ where $p_\theta(u) \geq \alpha$ and $\lambda$ represents the network learning rate. In $L_{SL}$, the default score for hard-labeled data is one. The concept is somewhat resembles to curriculum learning [2]. The architectural detailed is discussed in the following subsection.

### 3.2. The maskedFaceNet Model

As mentioned above, the proposed maskedFaceNet is inspired by a feed-forward SSD approach which is a collection of fixed size bounding boxes and scores that are produced for every possible object class instances. These predictions are then passed through non-maximum suppression (NMS) layer to compute the final detections. Our maskedFaceNet structure is based on the standard VGG-16 [40] like network with a few additional assistant layers. The VGG-16 like RFBNet [27] is truncated before the classification layers. The architecture till $conv_{5,x}$ is then followed by assistant layers, as shown in Figure 2.

Homogeneous to SSD architecture, maskedFaceNet is also a multi-scale one-stage framework for masked face detection. As shown in Figure 2, the fully connected ($fc$) layers are transformed to convolutional layers to reduce the complexity and the $maxpool$ layers are replaced with convolutional layers with stride two ($r = 2$) to learn the important properties while down sampling the input. The layers are decreased in size progressively which introduces multi-scale feature maps to detect different size faces. This makes the proposed model lightweight yet is powerful to capture the true face features with and without mask. Here, in this mode all additional assistant layers are randomly initialized with the "Xavier" initialization method [16].

Note, in the proposed maskedFaceNet model, $conv_{4}$, $conv_{5}$, $conv_{6}$ and $conv_{7}$ are used as the detection feature maps which associates to different anchor scales to predict distinct face sizes in an image. $BatchNorm$ layer is used after every $conv$ layers followed by $ReLU$, as shown in Figure 2. The assistant layer here is dedicated to the task of masked face detection and regulates feature maps accordingly.

The second last layer is the prediction layer, before multi-task loss layers, which is a $(u \times 3 \times 3 \times v)$ $conv$ layer where variable u and v denote the input and output channel number, respectively. The anchor output is a set of four offsets, related to bounding box coordinates and $N_c$ scores for...
classification, where \( c = 3 \) in our case. The anchors used in this paper is set to 1:1 aspect ratio, as the face bounding boxes mostly fit in a square quadrangle (approx). Followed by multi-task smooth\( L_1 \) loss and softmax loss layers for masked face bounding box regression and classification, respectively. The training strategy used to train the maskedFaceNet is discussed in detail in the very next subsection.

### 3.3. Training Strategy

As shown in Figure 3, the training of maskedFaceNet is done in three-phase fashion. In first phase, due to insufficient masked face image dataset, the proposed network is first trained on big WIDER face dataset [49] to initialize maskedFaceNet with proper weights to reduce the false positive cases in face detection. For this, we adopted grid loss [53] to learn the detector to detect even partial faces correctly. This learning helps further in detecting faces covered via masks or similar occlusions, which is our ultimate objective. We then fine-tune it on the proposed masked face datasets \( D: \text{MASK-face}_{vi} \), where \( i = \{1, 2\} \) (see Table 1 for datasets detail). Before the network is trained on masked dataset, we fixed the first two layers of maskedFaceNet, i.e., \( \text{conv1} \) and \( \text{conv2} \), to inherit the information learned from WIDER face dataset. Note, now the learning loss function for masked face detection is updated to smooth\( L_1 \) loss, due to its better consistence performance in object detection (see Section 4).

In second phase, the trained maskedFaceNet model is used to generate pseudo-labels for the unlabeled video frames related to COVID-19 collected from various local news channels over the Internet. Let’s say, the unlabeled video frame dataset \( U \) is semi-labeled and based on its confidence score \( \alpha \), a new pseudo-labeled dataset \( V \) is obtained, which is a subset of \( U \) and \( D < V \). It is then used to retrain maskedFaceNet parameter. As we know, the larger the dataset the better the learning is. Therefore, we believed this pseudo-labeled masked face dataset \( V \) will set a better network performance, similar to [48]. To make sure learning is not biased, the video frames are collected from difference sources and are of different resolutions. There are total ten video clips collected for experiment purpose of different lengths, counting to a total of 61,937 unlabeled images in \( U \) dataset. The generated pseudo-labeled dataset \( V \) with \( \alpha = 0.9 \) has 42,000 images, which is huge enough compared to MASK-face\(_{vi}\) dataset. This pseudo-labeled dataset is much suitable to train maskedFaceNet from scratch. The experimental details are discussed in Section 4. Note, if \( \alpha \) is varied, the performance analysis will differ. The size of \( V \) dataset can also be explored and can be set to a balanced class dataset. In this paper, we ignore this setting to strictly focus on the our proposed hypothesis of boosting the performance via progressive semi-supervised data.

In third phase, the semi-supervised trained maskedFaceNet model is fine-tuned again with the hard-labeled MASK-face\(_{vi}\) dataset to obtain the best masked face detection model with less labeled data. The flow of training phases are detailed in Figure 3. While training in phase-three, we introduce a progressive training for pseudo-labeled data. That is, the pseudo-labels are gradually pumped-in to the training process in such a way that the higher score are first entered followed by the lower scores. This actually updates the gradient based hyper-parameters which is based on the average over mini-batch size \( b \) of the complete training set, say \( V_N \). The noise scale \( \frac{\lambda_{\alpha}}{T} \) will keep network active, where \( \alpha \) ranges from 1 to \( V_N \). This approach will reduces the total computation cost yet reach to the optimal solution.

We also introduced a lighter version of maskedFaceNet that follows the same architecture with less convolutional layers, resulting a faster performing maskedFaceNet with almost equivalent results. That is to say, in the third phase a new lighter weighted maskedFaceNet with \( \Theta^* \) parameter which is \( < \Theta \) is also examined (see the experiment Section).

### 3.4. Loss Function

The overall objective of maskedFaceNet detector is to detect masked and non-masked faces in real-time with high accuracy rate. Therefore, similar to [34], the ultimate objective loss function \( \mathcal{L} \) of maskedFaceNet is a summation of regression loss \( \mathcal{L}_{reg} \) and classification loss \( \mathcal{L}_{cls} \) which is defined as:

\[
\mathcal{L} (p_c, v, v^*) = \frac{1}{N} (\mathcal{L}_{cls} (p_c) + \Lambda \times \mathcal{L}_{reg} (v, v^*))
\]

where \( N \) is the count of matched bounding boxes and \( \Lambda \) is a constant to balance these two terms which is set to 1 by default. Variable \( p_c \) is the corresponding probability and \( c \) represents the class. The set \( (v, v^*) \) is the predicted and ground truth (GT) bounding boxes for the corresponding class, respectively. The classification loss is a softmax loss over multiple classes \( c \), defined as:

\[
\mathcal{L}_{cls} (p_c) = - \log p_c
\]

and the regression loss \( \mathcal{L}_{reg} \) is defined as:

\[
\mathcal{L}_{reg} (v, v^*) = \sum_{i \in \{x, y, w, h\}} \left\{ \gamma (v_i - v_{i^*})^2 \quad \text{if} \ |v - v^*| < 1 \right\} \left\{ |v_i - v_{i^*}| - \gamma \quad \text{otherwise} \right\}
\]

where \( v \) an \( v^* \) is a four tuple vector with top left corner and width and height, i.e., \((v_x, v_y, v_w, v_h)\). \( \gamma = 0.5 \) and \( \mathcal{S} \) denotes the summation. And for final detection, we used smooth\( L_1 \) loss as it is less sensitive to outliers.
Table 1: Dataset analysis. *We split it as it was not provided by the author.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>#train</th>
<th>#test</th>
<th>GT Format</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>FaceMaskDataset</td>
<td>6132</td>
<td>1839</td>
<td>Pascal VOC</td>
<td>Detection</td>
</tr>
<tr>
<td>AfricanMaskedFaces</td>
<td>557</td>
<td>1389</td>
<td>Pascal VOC</td>
<td>Detection</td>
</tr>
<tr>
<td>MASK-face_v1</td>
<td>1689</td>
<td>198</td>
<td>Pascal VOC</td>
<td>Detection</td>
</tr>
<tr>
<td>MASK-face_v2</td>
<td>3142</td>
<td>335</td>
<td>Pascal VOC</td>
<td>Detection</td>
</tr>
</tbody>
</table>

Table 2: Comparison of maskedFaceNet on MASK-face_v1 and MASK-face_v2 datasets with various state-of-the-art methods.

<table>
<thead>
<tr>
<th>Methods</th>
<th>MASK-face_v1</th>
<th>MASK-face_v2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN ResNet-50</td>
<td>0.640</td>
<td>0.780</td>
</tr>
<tr>
<td>SSD300 [28]</td>
<td>0.470</td>
<td>0.510</td>
</tr>
<tr>
<td>FaceBoxes [53]</td>
<td>0.920</td>
<td>0.910</td>
</tr>
<tr>
<td>SSD_MobileNet [28]</td>
<td>0.490</td>
<td>0.690</td>
</tr>
<tr>
<td>maskedFaceNet</td>
<td>0.977</td>
<td>0.981</td>
</tr>
<tr>
<td>maskedFaceNet light</td>
<td>0.976</td>
<td>0.981</td>
</tr>
</tbody>
</table>

Table 3: Comparison of state-of-the-art methods along with the proposed models on FaceMaskDataset [8].

<table>
<thead>
<tr>
<th>Methods</th>
<th>FaceMaskDataset</th>
<th>Pre-trained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline [8]</td>
<td>0.9075</td>
<td>ImageNet</td>
</tr>
<tr>
<td>RetinaMask w MobileNet</td>
<td>0.8165</td>
<td>ImageNet</td>
</tr>
<tr>
<td>RetinaMask w ResNet</td>
<td>0.9210</td>
<td>ImageNet</td>
</tr>
<tr>
<td>RetinaMask w MobileNet</td>
<td>0.8265</td>
<td>WIDER</td>
</tr>
<tr>
<td>RetinaMask w ResNet</td>
<td>0.9265</td>
<td>WIDER</td>
</tr>
<tr>
<td>maskedFaceNet</td>
<td>0.9554</td>
<td>WIDER</td>
</tr>
<tr>
<td>maskedFaceNet light</td>
<td>0.9405</td>
<td>WIDER</td>
</tr>
</tbody>
</table>

3.5. Implementation Detail

For implementation, maskedFaceNet uses WIDER face dataset to initialize the weights using grid loss. The weights
are then updated using learning rate $\lambda = 0.0001$, in the first phase, to fine-tune the model on MASK-face_v1 dataset for 50 epochs. In the second phase, $\lambda$ is set to 0.001 and trained for another 50 epochs on pseudo-labeled dataset $V$. In the third phase, the model is fine-tuned with $\lambda = 0.00001$ for 30 epochs and decay of 0.1 at every 10 epochs. For all the three phases, the weight decay $\omega$ and momentum $\mu$ are set to $5 \times 10^{-2}$ and 0.9, respectively.

The experiments are all conducted on Intel Xeon workstation with NVIDIA Titan X 12GB GPU on Pytorch4 platform in Ubuntu 18.04 environment.

4. Experiments and Results

In this section, we firstly introduce the two established masked face datasets to be public for researchers to analyze the possible research directions in this field. In Table 1 both the datasets distribution are detailed and some sample images are shown in Figure 4. We also used a recently published FaceMaskDataset [8] and African Masked Face Dataset (Figure 1f) for experiments and further comparison with the state-of-the-art methods. Followed by various different combinations of parameter settings to validate our light weighted architecture for the real-time masked face detection.

In this paper, to evaluate the performance of the proposed model and other state-of-the-art methods, we used mean average precision (mAP).

Table 4: Comparison of maskedFaceNet on MASK-face_v1 and MASK-face_v2 datasets with smoothL1-loss and different resolution.

<table>
<thead>
<tr>
<th>MASK-face_v1 Dataset</th>
<th>Method</th>
<th>mAP (320)</th>
<th>mAP (640)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>maskedFaceNet</td>
<td>0.9765</td>
<td>0.9775</td>
</tr>
<tr>
<td></td>
<td>maskedFaceNet light</td>
<td>0.9753</td>
<td>0.9768</td>
</tr>
<tr>
<td>MASK-face_v2 Dataset</td>
<td>Method</td>
<td>mAP (320)</td>
<td>mAP (640)</td>
</tr>
<tr>
<td></td>
<td>maskedFaceNet</td>
<td>0.9810</td>
<td>0.9813</td>
</tr>
<tr>
<td></td>
<td>maskedFaceNet light</td>
<td>0.9800</td>
<td>0.9812</td>
</tr>
</tbody>
</table>

4.1. Datasets

As mentioned above, since there is no such publicly available wild dataset for mixed masked faces for face detection, we introduce two new masked face datasets called MASK-face_v1 and MASK-face_v2. The datasets consists of real natural scenes of populations from indoor and outdoor. To make balance between masked and non-masked faces, we equally inherited face images from WIDER face dataset [49]. In the first version of our dataset, the annotation includes ‘face’ and ‘mask’ with a bounding box covering the complete head while in the second version, only faces are bounded, can see the difference in Figure 4. In addition, we also compare our proposed method with a recently published FaceMaskDataset [8] which is a subset of MAFA [14] and WIDER face datasets. Lastly, we merged our MASK-face_v2 with African Masked Face dataset to let network learn different color and shape of masks on different skin color faces (see the ablation study for experimental results). Since there is no separate training and testing set, we split African Masked Face dataset into 80%-20% ratio.

Figure 5: Semi-supervised masked face annotation via maskedFaceNet. First row: MASK-face_v1 head annotated and MASK-face_v2 face annotated. Classes are color coded.

4.2. Model Analysis

4.2.1 Comparison with state-of-the-art methods

We implemented and compared various state-of-the-art ODs and FDs on the public dataset and also on our proposed datasets. Table 2 shows a detailed comparison on our established datasets with other state-of-the-art methods. In Table 3 we compared maskedFaceNet with various versions of RetinaMask [18] FD on FaceMaskDataset [8]. It is observed that the proposed model performance is improved by 4% and 6% compared to the second best method, i.e., FaceBoxes [53] on MASK-face_v1 and MASK-face_v2 datasets, respectively. While in cases of FaceMaskDataset, RetinaNet with ResNet backbone manages to secure the second best position with a mean precision rate of 92.65% where our maskedFaceNet achieves the best mAP.

4.2.2 Comparisons with different settings

In the second set of experiments, we compare the performance index of the proposed method with different input resolutions, different objective loss functions and different batch sizes. Table 4 shows comparison of maskedFaceNet with 320 and 640 input image frames where the performance is observed to be quite similar. In Table 5 three different objective loss functions are analyzed to conclude that smoothL1 loss is more suitable for mask face detection.

4.3. Conclusion

In this paper, we present a novel masked face detection method called maskedFaceNet which uses two stages multi-detect architecture for face and mask detection. The experiments are all conducted on Intel Xeon workstation with NVIDIA Titan X 12GB GPU on Pytorch4 platform in Ubuntu 18.04 environment.

Appendix A

A.1. Implementation details

The experiments are conducted on Intel Xeon workstation with NVIDIA Titan X 12GB GPU on Pytorch4 platform in Ubuntu 18.04 environment. The network is implemented using Pytorch and CUDA. The training is done using 64GB memory and 500000 iterations. The learning rate is set to 0.0001 for the first phase, 0.001 for the second phase and 0.00001 for the third phase. The weight decay is set to 5 \times 10^{-2} and momentum is set to 0.9 for all phases.

A.2. Ablation Study

The ablation study shows the importance of each component of the proposed method. The results show that the use of smoothL1 loss function, multi-detection architecture and semi-supervised training significantly improve the performance of masked face detection.
Table 5: mAP comparison of maskedFaceNet on MASK-face \textsubscript{v2} with different loss functions and batch sizes.

<table>
<thead>
<tr>
<th>Methods</th>
<th>L1-loss</th>
<th>Wing-loss</th>
<th>SmoothL1-loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>maskedFaceNet</td>
<td>0.9779</td>
<td>0.9774</td>
<td>0.9813</td>
</tr>
<tr>
<td>maskedFaceNet light</td>
<td>0.9750</td>
<td>0.9751</td>
<td>0.9812</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MASK-face \textsubscript{v2} Dataset</th>
<th>Batch Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b=8</td>
</tr>
<tr>
<td>maskedFaceNet</td>
<td>0.9789</td>
</tr>
<tr>
<td>maskedFaceNet light</td>
<td>0.9782</td>
</tr>
</tbody>
</table>

Table 6: mAP comparison of maskedFaceNet on MASK-face \textsubscript{v2} with and without semi-supervised data.

<table>
<thead>
<tr>
<th>Methods</th>
<th>semi-supervised</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>maskedFaceNet</td>
<td>\times</td>
<td>0.9533</td>
</tr>
<tr>
<td>maskedFaceNet light</td>
<td>\checkmark</td>
<td>0.9813</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>MASK-face \textsubscript{v2} + African Masked Face Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>maskedFaceNet</td>
<td>\times</td>
</tr>
<tr>
<td>maskedFaceNet light</td>
<td>\checkmark</td>
</tr>
</tbody>
</table>

5. Conclusions and Future Direction

In this paper, we have proposed a new light weighted maskedFaceNet for real-time masked face detection. The proposed model gets benefit from progressive semi-supervised learning which focus on pseudo-labels that are generated \textit{via} the initial stage of the model itself to obtain a better weight initialization. We also explored the suitable objective loss function for masked FDs. For the study, we established two different real wild masked face datasets. The experimental results on different datasets show that the proposed maskedFaceNet outperforms the current state-of-the-art methods and indicates the effectiveness of the proposed hypothesis for all types of datasets used in this paper.

In further, the work will be directed towards masked face recognition along with incorrect masked face detection. And also will try to implement this progressive semi-supervised hypothesis on other challenging tasks where data size is limited and data annotation is challenging such as industrial applications. Another dimension is domain adaptation \textit{via} semi-supervised training learning.

References


