maskedFaceNet: A Progressive Semi-Supervised Masked Face Detector

Shitala Prasad, Yiqun Li, Dongyun Lin, Dong Sheng
Visual Intelligence, Institute for Infocomm Research (I^2R), 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

(a) (b) (c) (d) (e) (f)

Figure 1: Masked face detection challenges. (a) Single face with and without mask, (b) Masked and non-masked faces in same image wearing different color masks, (c) Side-by-faces with and without mask, (d) Occluded faces with mask or non-mask objects, (e) Different color mask and PPE and (f) Dark faces with and without mask. Note: the images are collected from various local news channels available over Internet.

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 handcrafted 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, 36] 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 [4] while Ghiasi and Fowlkes proposed a joint detector that can handle face detection as well as key feature localization [15]. 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.* [3] proposed STN with a cascade CNN to address the challenge of huge face pose variation in real-word 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 performs 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 MAsked 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[1] 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[2] 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 $\mathcal{D} = \{(x, y)\}$, and an unlabeled image dataset, say $\mathcal{U} = \{(u)\}$, SSL aims to solve the following problem:

$$\min_{\theta} \sum_{(x,y)\epsilon\mathcal{D}} \mathcal{L}_{SL}(x, y, \theta) + \beta \sum_{(u)\epsilon\mathcal{U}} \mathcal{L}_{UL}(u, \theta) \quad (1)$$

where $\mathcal{L}_{SL}$ and $\mathcal{L}_{UL}$ represents the supervised and unlabeled loss, respectively. The value θ is the total trainable parameters of the given model and β is the weight balancing parameter which is $\mathbf{R}_{>0}$. 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 $\mathcal{D} << \mathcal{U}$. There are different ways to compute pseudo-labels for $u\epsilon\mathcal{U}$ and calculate the per-example unsupervised or semi-supervised loss ($\mathcal{L}_{UL}$) 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 $\mathcal{L}_{UL}$. This improves the performance of SSL algorithm but leads to the computational expense and there-

[1]France is using AI to detect whether people are wearing masks https://slate.com/technology/2020/05/france-artificial-intelligence-mask-detection-coronavirus.html

[2]https://www.nist.gov/news-events/news/2020/07/nist-launches-studies-masks-effect-face-recognition-software

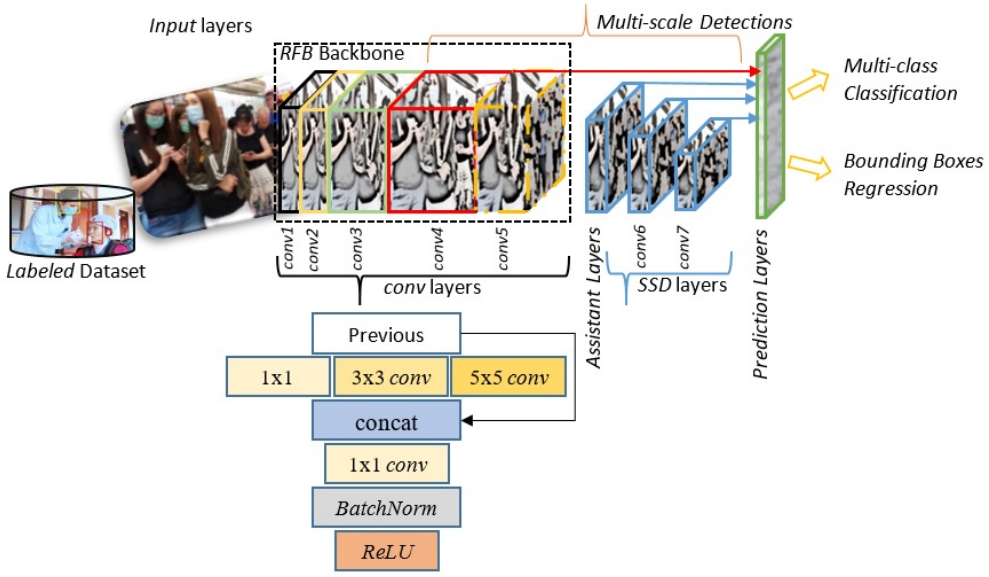

Figure 2: Architecture of maskedFaceNet consisting of convolutional layers, *BatchNorm* and *ReLU* along with dedicated assistant layers, prediction layers and multi-task loss layers (color coded for convenience).

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 α for all $u \epsilon \mathcal{U}$. Specifically, instead of obtaining only the pseudo-label [20] 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:

$$\min_{\Theta} \sum_{(X,Y)\epsilon\mathcal{D}} \mathcal{L}_{SL}(X, Y, \Theta) + \beta \sum_{(U,\tilde{Y})\epsilon\mathcal{U}} \mathcal{L}_{UL}(U, \tilde{Y}, \Theta) \tag{2}$$

where (X, Y) and $(U, \tilde{Y})$ 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 $\tilde{Y}$ there is an extra score tuple α for fair Θ parameters learning. Thus, the weight importance $\mathbf{w}_u$ for the unlabeled data is initialized as $\lambda * \alpha$ where $p_\theta(u) \geq \alpha$ and λ represents the network learning rate. In $\mathcal{L}_{SL}$, 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 *conv5_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, *conv4*, *conv5*, *conv6* and *conv7* 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 $(\mathbf{u} \times 3 \times 3 \times \mathbf{v})$ *conv* layer where variable $\mathbf{u}$ and $\mathbf{v}$ 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 *smoothL1* 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 [33] 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 $\mathcal{D}$: 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.*, *conv1* and *conv2*, to inherit the information learned from WIDER face dataset. Note, now the learning loss function for masked face detection is updated to *smoothL1* 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 $\mathcal{U}$ is semi-labeled and based on its confidence score α, a new pseudo-labeled dataset $\mathcal{V}$ is obtained, which is a subset of $\mathcal{U}$ and $\mathcal{D} < \mathcal{V}$. 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 $\mathcal{V}$ 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 $\mathcal{U}$ dataset. The generated pseudo-labeled dataset $\mathcal{V}$ 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 α is varied, the performance analysis will differ. The size of $\mathcal{V}$ 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 $\mathcal{V}_N$. The noise scale $\frac{\lambda \times \mathcal{V}_i}{b}$ will keep network active, where i ranges from 1 to $\mathcal{V}_N$. 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 Θ' 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:

$$\mathcal{L}\left(p_c, v, v*\right) = \frac{1}{N}(\mathcal{L}_{cls}\left(p_c\right) + \Lambda \times \mathcal{L}_{reg}\left(v, v*\right) \qquad (3)$$

where N is the count of matched bounding boxes and Λ 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:

$$\mathcal{L}_{cls}\left(p_c\right) = -\log p_c \qquad (4)$$

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

$$\mathcal{L}_{reg}\left(v, v*\right) = \mathcal{S}_{i\ \epsilon\ \{x,y,w,h\}} \begin{cases} \gamma\left(v_i - v_i*\right)^2 & \text{if } |v - v*| < 1 \\ |v_i - v_i*| - \gamma & \text{otherwise} \end{cases} \qquad (5)$$

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 *smoothL1* loss as it is less sensitive to outliers.

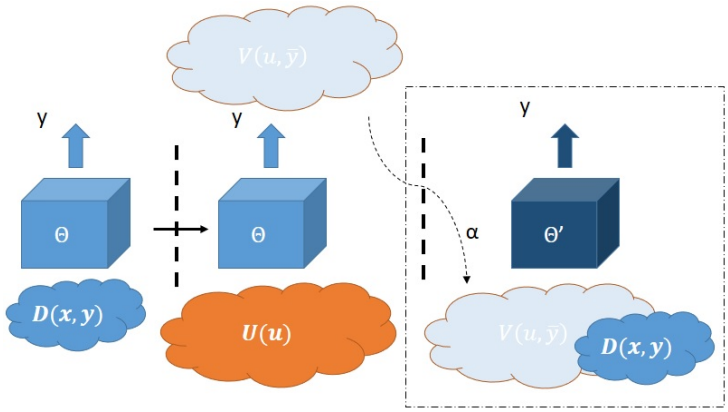

Figure 3: The three-phase SSL training scheme.

Table 1: Dataset analysis. *We split it as it was not provided by the author.

Datasets	#train	#test	GT Format	Task
FaceMaskDataset [8]	6132	1839	Pascal VOC	Detection
AfricanMaskedFaces[3]	5557*	1389*	Pascal VOC	Detection
MASK-face_*v1*	1689	198	Pascal VOC	Detection
MASK-face_*v2*	3142	335	Pascal VOC	Detection

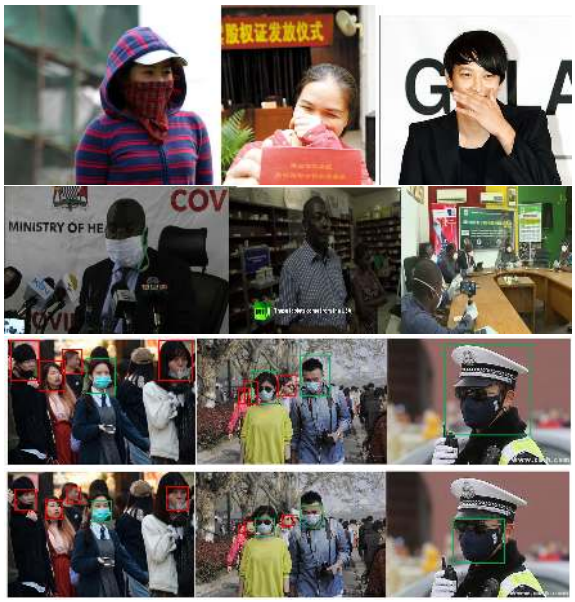

Figure 4: Annotated masked face datasets. First row: FaceMaskDataset, Second row: AfricanMaskedFaces dataset Third row: MASK-face_*v1* head annotated dataset and Fourth row: MASK-face_*v2* face annotated dataset. Classes are color coded.

Table 2: Comparison of maskedFaceNet on MASK-face_*v1* and MASK-face_*v2* datasets with various state-of-the-art methods.

Methods	mAP	
	MASK-face_*v1*	MASK-face_*v2*
Faster R-CNN ResNet-50 [37]	0.640	0.780
SSD300 [28]	0.470	0.510
FaceBoxes [53]	0.920	0.910
SSD_MobileNet [28]	0.490	0.690
maskedFaceNet	**0.977**	**0.981**
maskedFaceNet *light*	0.976	0.981

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

Methods	FaceMaskDataset	
	mAP	Pre-trained
Baseline [8]	0.9075	ImageNet
RetinaMask *w* MobileNet [18]	0.8165	ImageNet
RetinaMask *w* ResNet [18]	0.9210	ImageNet
RetinaMask *w* MobileNet [18]	0.8265	WIDER
RetinaMask *w* ResNet [18]	0.9265	WIDER
maskedFaceNet	**0.9554**	WIDER
maskedFaceNet light	0.9405	WIDER

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_*vi* dataset for 50 epochs. In the second phase, λ is set to 0.001 and trained for another 50 epochs on pseudo-labeled dataset $\mathcal{V}$. 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 ω and momentum μ are set to 5×10^{-2} and 0.9, respectively.

The experiments are all conducted on Intel Xeon workstation with NVIDIA Titan X 12GB GPU on Pytorch[4] 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 FaceMaskedDataset [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.

MASK-face_*v1* Dataset		
Method	**mAP (320)**	**mAP (640)**
maskedFaceNet	**0.9765**	**0.9775**
maskedFaceNet light	0.9753	0.9768

MASK-face_*v2* Dataset		
Method	**mAP (320)**	**mAP (640)**
maskedFaceNet	**0.9810**	**0.9813**
maskedFaceNet light	0.9800	0.9812

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.

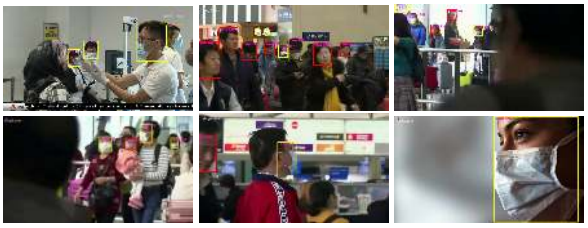

Figure 5: Semi-supervised masked face annotation *via* maskedFaceNet. First row: MASK-face_*v1* head annotated and MASK-face_*v2* face annotated. Classes are color coded.

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]Pytorch: `https://pytorch.org/`

Table 5: mAP comparison of maskedFaceNet on MASK-face_*v2* with different loss functions and batch sizes.

MASK-face_*v2* Dataset									
Methods	***L1-loss***	***Wing-loss***	***SmoothL1-loss***	**Batch Size**					
				b=8	***b=12***	***b=16***	***b=24***	***b=48***	***b=72***
maskedFaceNet	0.9779	0.9774	**0.9813**	0.9789	0.9782	0.9789	**0.9813**	0.9758	0.9777
maskedFaceNet light	0.9750	0.9751	**0.9812**	0.9782	0.9782	09781	**0.9812**	0.9746	0.9768

Table 6: mAP comparison of maskedFaceNet on MASK-face_*v2* with and without semi-supervised data.

Methods	**MASK-face_*v2* Dataset**	
	semi-supervised	**mAP**
maskedFaceNet	×	0.9533
maskedFaceNet	√	**0.9813**
maskedFaceNet *light*	×	0.9480
maskedFaceNet *light*	√	**0.9812**

Methods	**MASK-face_*v2* + African Masked Face Dataset**
maskedFaceNet	×	0.9636
maskedFaceNet *w/o* progressive	√	0.9756
maskedFaceNet	√	0.9811

task. Second half of Table 5 shows the role of batch size while training maskedFaceNet on masked face dataset and found that batch size 24 performances the best even when the training dataset is small.

4.2.3 Ablation Study

To evaluate the contribution of the proposed technique that progressive semi-supervised dataset plays a major role in training masked FD when the training data itself is small enough, an ablation study is performed. Table 6 shows comparison between with and without semi-supervised training dataset and is observed that if the model is pre-trained on huge semi-supervised data, which is $\sim$13$\times$ of MASK-face_*v2*, and then fine-tuned in progressive semi-supervised fashion, as shown in Figure 3, the performance boosts by at least 2% to 3.3%. This improvement depends upon the number of pseudo labels $\mathcal{V}_N$ generated *via* maskedFaceNet and their score threshold α. A recursive process of re-generation of semi-supervised data $\mathcal{V}$ and re-training maskedFaceNet will further improve the accuracy but due to the scope of the proposed hypothesis, we limit it to one recursion only. Figure 5 shows some of the sample images with their pseudo-labels $(U, \tilde{Y}, \alpha)$ that are generated automatically after training on MASK-face_*v1* and/or MASK-face_*v2* datasets.

The second experiment we performed is by mixing different skin color faces, *i.e.*, MASK-face_*v2* + African Masked Face datasets. The results are listed in the lower part of Table 6. The role of the proposed approach is clearly differentiable from [20] like approaches.

Finally, we calculated the compute power of the proposed maskedFaceNet on both GPU and CPU, for the real-time analysis. It is observed that maskedFaceNet is well suitable for CPU computation to achieve the real-time scenario with 38fps, see in Table 7.

Table 7: Computation comparison of maskedFaceNet on MASK-face_*v2* for 640 input resolution.

MASK-face_*v2* Dataset			
Methods	**Computation**		**#params**
maskedFaceNet	FPS-GPU	37	568,348
	FPS-CPU	29	
maskedFaceNet light	FPS-GPU	44	536,140
	FPS-CPU	38	

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 *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 *via* semi-supervised training learning.

References

[1] Mauro Baquero-Suárez, John Cortés-Romero, Jaime Arcos-Legarda, and Horacio Coral-Enriquez. A robust two-stage active disturbance rejection control for the stabilization of a riderless bicycle. *Multibody System Dynamics*, 45(1):7–35, 2019.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[3] Dong Chen, Gang Hua, Fang Wen, and Jian Sun. Supervised transformer network for efficient face detection. In *ECCV*, pages 122–138. Springer, 2016.

[4] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *ECCV*, pages 109–122. Springer, 2014.

[5] Weijun Chen, Hongbo Huang, Shuai Peng, Changsheng Zhou, and Cuiping Zhang. Yolo-face: a real-time face detector. *The Visual Computer*, pages 1–9, 2020.

[6] Yujia Chen, Lingxiao Song, Yibo Hu, and Ran He. Adversarial occlusion-aware face detection. In *BTAS*, pages 1–9. IEEE, 2018.

[7] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou. Selective refinement network for high performance face detection. In *AAAI*, volume 33, pages 8231–8238, 2019.

[8] Daniell Chiang. Detect faces and determine whether people are wearing mask, 2020 (accessed May 15, 2020).

[9] Daniel Crouse, John E Bradley III, and Lewis C Lee. Aggregating procedures for automatic document analysis, Jan. 7 2020. US Patent 10,528,609.

[10] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019.

[13] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Trans. PAMI*, 32(9):1627–1645, 2009.

[14] Shiming Ge, Jia Li, Qiting Ye, and Zhao Luo. Detecting masked faces in the wild with lle-cnns. In *CVPR*, pages 2682–2690, 2017.

[15] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *CVPR*, pages 2385–2392, 2014.

[16] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[17] Youngkyoon Jang, Hatice Gunes, and Ioannis Patras. Registration-free face-ssd: Single shot analysis of smiles, facial attributes, and affect in the wild. *Computer Vision and Image Understanding*, 182:17–29, 2019.

[18] Mingjie Jiang and Xinqi Fan. Retinamask: A face mask detector. *arXiv preprint arXiv:2005.03950*, 2020.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[20] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013.

[21] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *CVPR*, pages 5325–5334, 2015.

[22] Jianguo Li and Yimin Zhang. Learning surf cascade for fast and accurate object detection. In *CVPR*, pages 3468–3475, 2013.

[23] Shengcai Liao, Anil K Jain, and Stan Z Li. A fast and accurate unconstrained face detector. *Trans. PAMI*, 38(2):211–223, 2015.

[24] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *ICIP*, volume 1, pages I–I. IEEE, 2002.

[25] Shaohui Lin, Ling Cai, Xianming Lin, and Rongrong Ji. Masked face detection via a modified lenet. *Neurocomputing*, 218:197–202, 2016.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[27] Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *ECCV*, pages 385–400, 2018.

[28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

[29] Lada Maleš, Darijan Marčetić, and Slobodan Ribarić. A multi-agent dynamic system for robust multi-face tracking. *Expert Systems with Applications*, 126:246–264, 2019.

[30] D Manju and V Radha. A novel approach for pose invariant face recognition in surveillance videos. *Procedia Computer Science*, 167:890–899, 2020.

[31] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *Trans. PAMI*, 2020.

[32] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in neural information processing systems*, pages 3235–3246, 2018.

[33] Michael Opitz, Georg Waltner, Georg Poier, Horst Possegger, and Horst Bischof. Grid loss: Detecting occluded faces. In *ECCV*, pages 386–402. Springer, 2016.

[34] Shitala Prasad and Adams Wai Kin Kong. Using object information for spotting text. In *ECCV*, pages 540–557, 2018.

[35] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *Trans. PAMI*, 41(1):121–135, 2017.

[36] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[38] Zhongzheng Ren, Raymond A Yeh, and Alexander G Schwing. Not all unlabeled data are equal: Learning to weight data in semi-supervised learning. *arXiv preprint arXiv:2007.01293*, 2020.

[39] S Shishira, Vidyadhar Rao, and Sithu D Sudarsan. Proximity contours: Vision based detection and tracking of objects in manufacturing plants using industrial control systems. In *International Conference on Industrial Informatics*, volume 1, pages 1021–1026. IEEE, 2019.

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[41] Pankaj Pratap Singh, Shitala Prasad, Anil Kumar Chaudhary, Chandan Kumar Patel, and Manisha Debnath. Classification of effusion and cartilage erosion affects in osteoarthritis knee mri images using deep learning model. In *Computer Vision and Image Processing*, pages 373–383, Singapore, 2020. Springer Singapore.

[42] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[43] Xu Tang, Daniel K Du, Zeqiang He, and Jingtuo Liu. Pyramidbox: A context-assisted single shot face detector. In *ECCV*, pages 797–813, 2018.

[44] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, volume 1, pages I–I. IEEE, 2001.

[45] Tobias Vogelpohl, Matthias Kühn, Thomas Hummel, and Mark Vollrath. Asleep at the automated wheel—sleepiness and fatigue during highly automated driving. *Accident Analysis & Prevention*, 126:70–84, 2019.

[46] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.

[47] Shudong Xie, Yiqun Li, Dongyun Lin, Tin Lay Nwe, and Sheng Dong. Meta module generation for fast few-shot incremental learning. In *ICCV Workshops*, pages 1–10, 2019.

[48] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

[49] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, pages 5525–5533, 2016.

[50] Qiting Ye. Masked face detection via a novel framework. In *International Conference on Mechanical, Electronic, Control and Automation Engineering*. Atlantis Press, 2018.

[51] Jialiang Zhang, Xiongwei Wu, Steven CH Hoi, and Jianke Zhu. Feature agglomeration networks for single stage face detection. *Neurocomputing*, 380:180–189, 2020.

[52] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *Signal Processing Letters*, 23(10):1499–1503, 2016.

[53] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. Faceboxes: A cpu real-time face detector with high accuracy. In *International Joint Conference on Biometrics*, pages 1–9. IEEE, 2017.

[54] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *ICCV*, pages 192–201, 2017.

[55] Zhishuai Zhang, Wei Shen, Siyuan Qiao, Yan Wang, Bo Wang, and Alan Yuille. Robust face detection via learning small faces on hard images. In *WACV*, pages 1361–1370, 2020.

[56] Wan-Lei Zhao and Chong-Wah Ngo. Flip-invariant sift for copy and object detection. *Trans. IP*, 22(3):980–991, 2012.

[57] Chenchen Zhu, Ran Tao, Khoa Luu, and Marios Savvides. Seeing small faces from robust anchor's perspective. In *CVPR*, pages 5127–5136, 2018.

[58] Chenchen Zhu, Yutong Zheng, Khoa Luu, and Marios Savvides. Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection. In *Deep learning for biometrics*, pages 57–79. Springer, 2017.