Detecting Masked Faces in the Wild with LLE-CNNs

Shiming Ge[1], Jia Li[2*], Qiting Ye[1,3], Zhao Luo[1,3]

[1]Beijing Key Laboratory of IOT information Security, Institute of Information Engineering, Chinese Academy of Sciences

[2]State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University

[3]School of Cyber Security, University of Chinese Academy of Sciences

Abstract

*Detecting faces with occlusions is a challenging task due to two main reasons: 1) the absence of large datasets of masked faces, and 2) the absence of facial cues from the masked regions. To address these two issues, this paper first introduces a dataset, denoted as **MAFA**, with* $30,811$ *Internet images and* $35,806$ *masked faces. Faces in the dataset have various orientations and occlusion degrees, while at least one part of each face is occluded by mask. Based on this dataset, we further propose LLE-CNNs for masked face detection, which consist of three major modules. The Proposal module first combines two pre-trained CNNs to extract candidate facial regions from the input image and represent them with high dimensional descriptors. After that, the Embedding module is incorporated to turn such descriptors into a similarity-based descriptor by using locally linear embedding (LLE) algorithm and the dictionaries trained on a large pool of synthesized normal faces, masked faces and non-faces. In this manner, many missing facial cues can be largely recovered and the influences of noisy cues introduced by diversified masks can be greatly alleviated. Finally, the Verification module is incorporated to identify candidate facial regions and refine their positions by jointly performing the classification and regression tasks within a unified CNN. Experimental results on the **MAFA** dataset show that the proposed approach remarkably outperforms 6 state-of-the-arts by at least 15.6%.*

1. Introduction

With the rapid development of machine learning methods, the problem of face detection seems to be well addressed yet. For example, the face detector proposed in [17] achieves an average precision of 98.0% on the public image benchmark AFW [37] by using the cascaded Convolutional Neural Networks, while the speed of some face detectors can reach up to 35 FPS [1] or even 400 FPS [21].

* Corresponding author: Jia Li (email: jiali@buaa.edu.cn).

Figure 1. Masked faces may have diversified orientations, degrees of occlusion and mask types, making their detection an extremely challenging task for existing face detectors.

Due to the great success of these face detectors, some of them have been integrated into applications so as to facilitate auto-focusing [10], human computer interaction [12] and image database management [27].

Beyond the remarkable success achieved by existing works, there is increasing concern that the development of better face detectors is now becoming more and more difficult. In particular, the detection of masked faces, which can be very helpful for applications like video surveillance and event analysis, is still a challenging task for many existing models. As shown in Fig. 1, masked faces may have different orientations, degrees of occlusion and diversified types of masks, which make the accurate detection of masked faces a really challenging task even for the state-of-the-art face detectors [25, 4, 36, 19, 35]. Compared with the classic task of normal face detection, existing models often have a sharp performance drop in detecting masked faces, which may be mainly caused by two reasons. First, there lacks a large dataset with massive masked faces for exploring the key attributes shared by various masked faces and identifying the models with the state-of-the-art performance. Second, facial features from the occluded parts are no longer available in the detection process, while the existence of masks inevitably bring in certain kinds of noise. With *insufficient* training and testing data as well as *incomplete* and *inaccurate* features, masked face detection has been becoming a widely recognized challenging task in the area of face detection. Although this issue has been tentatively studied

in some recent works such as [7, 23, 32], it is still necessary to construct large datasets and develop effective and efficient models for masked face detection.

Toward this end, this paper presents a dataset for masked face detection, which is denoted as **MAFA**. The dataset consists of $30,811$ Internet images, in which $35,806$ masked human faces are manually annotated. In the annotation process, we ensure that each image contains at least one face occluded by various types of masks, while the six main attributes of each masked face, including locations of faces, eyes and masks, face orientation, occlusion degree and mask type, are manually annotated and cross-checked by nine subjects. The dataset will be released soon on the Internet, which we believe can facilitate the development of new face detectors in the future.

By inspecting the main characteristics of masked faces in **MAFA**, we find that most facial attributes can be lost in heavily occluded faces (*e.g.*, faces with only eyes unoccluded by masks), while the highly diversified masks can bring in various types of noises. Inspired by this fact, we propose LLE-CNNs for masked face detection by recovering missing facial cues and suppressing non-facial cues in the feature subspace. The proposed approach consists of a proposal module, an embedding module and a verification module. The proposal module first extracts a set of face proposals and characterize each proposal with a 4096d descriptor with a pre-trained VGG-Face model [24]. Considering that the descriptor of a masked face can be incomplete or noisy, we further embed it into a feature subspace formed by two dictionaries that consist of the descriptors from representative normal faces and non-faces. Note that such dictionaries are learned from a large pool of normal faces, masked faces and non-faces from previous datasets [14, 32] and the training set of **MAFA**. With an approximate locally linear embedding, a candidate region can be characterized by the similarity scores to representative normal faces and non-faces. Finally, such similarity-based descriptor is fed into the Verification module that consists of a Deep Neural Networks with only Fully-Connected (`FC`) layers so as to identify the real faces. Experimental results on the proposed **MAFA** dataset show that the proposed LLE-CNNs significantly outperform 6 state-of-the-arts [18, 20, 37, 22, 7, 35] in detecting masked faces.

The main contributions of this paper are three folds. 1) We present a dataset of masked faces that can be used as an additional training source for developing new face detectors; 2) We propose LLE-CNNs for masked face detection, which outperforms 6 state-of-the-art face detectors in detecting masked faces; and 3) We conduct a comprehensive analysis on the key challenges in masked face detection, which may be helpful for developing new face detectors in the future.

2. Related Work

Since the main contributions of this work include a new dataset and a novel model for masked face detection, we will first review the most popular datasets in the literature, followed by a briefly survey the state-of-the-art face detection models.

2.1. Datasets for Face Detection

In the literature, many datasets have been constructed to assess face detection models. While early datasets mainly consist of images collected in the controlled environment, recent datasets like MALF [30], IJB-A [13],CelebA [32] and WIDER FACE [31] tend to collect images from Internet. In this manner, these datasets demonstrate better capability in revealing the actual performances of face detectors in the wild. Moreover, recent face datasets often provide more annotations than early datasets. For example, in FDDB [11] only the position of each face is annotated. In contrast, the latest datasets such as OFD-UCI [7], MALF [30] and WIDER FACE [31] provide multiple facial attributes like positions, landmarks, gender, and poses.

In particular, compared with early datasets like FDDB [11], recent datasets are becoming much larger and contain much more real-world scenarios. For example, CelebA [32] is 71 times larger than FDDB in number of images, while the number of faces in WIDER FACE [31] is 76 times larger than that in FDDB. The rapidly growing face datasets constructs a more realistic testing platform for face detection models so that their application in real-world applications becomes much easier. Moreover, the presence of such large-scale datasets enable the usage of deep learning algorithms [15] that are extremely powerful in learning effective face representations directly from data.

2.2. Models for Face detection

As stated in [33], existing face detection models can be roughly grouped into three categories, in which models are based on boosting, Deformable Part Model (DPM) [6] and Convolutional Neural Network (CNN) [16], respectively.

1) Boosting-based category. In this category, the Viola-Jones face detector [28] is the one of the most famous models which adopted the boosted cascade with simple Haar features. Inspired by this model, Li *et al.* [18] proposed a multi-view face detector that adopted the surf features in the training and testing processes. In [21], a face detector was proposed to efficiently detect faces with an ensemble of optimized decision trees. By comparing pixel intensities in the internal nodes, faces can be detected at an extremely high speed. Recently, Liao *et al.* [20] proposed a face detector that utilized the scale invariant and bounded image feature called Normalized Pixel Difference. A single soft-cascade classifier was then adopted for efficient face detection. Gen-

erally speaking, face detectors in the boosting-based category are often very efficient.

2) DPM-based category. Beyond boosting-based methods, some approaches propose to explicitly model the structure or deformation of faces with DPM. For example, Zhu and Ramanan [37] proposed a tree structured model for face detection, which can simultaneously estimate face poses and localize facial landmarks. Mathias *et al.* [22] trained a DPM-based face detector with about $26,000$ faces from AFLW, which achieved an average precision of 97.14% on AFW [37]. By observing that aligned face shapes can provide better features for face classification, Chen *et al.* [1] presented a face detector by jointly learning detection and alignment in a unified framework. In [7], a model was proposed to jointly handle face detection and keypoint localization by using hierarchical DPM. Typically, DPM-based face detectors achieve impressive accuracy but may suffer from the high computational cost due to the usage of DPM.

3) CNN-based category. Different from the boosting-based and DPM-based approaches, CNN-based face detectors directly learn face representations from data [34, 5] and adopt deep learning paradigm [9, 8, 26] to detect the presence of a face in a scanning window. For example, Li *et al.* [17] proposed CascadeCNN, which is a boosted exemplar-based face detector. In [5], Farfade *et al.* fine-tuned the AlexNet [15] to get a multi-view face detector, which was trained on $200,000$ positive samples and 20 million negative samples. Yang *et al.* [29] proposed a face detector that utilized the feature aggregation framework [3], while the features were generated via CNN. In [32], the faceness of a window was assessed with an attribute-aware CNN, and occlusions were considered as well in generating face proposals. As a result, this method demonstrates strong capability in detecting faces with severe occlusion and pose variation.

More recently, Zhu *et al.* [36] presented Contextual Multi-Scale Region-based Convolution Neural Network (CMS-RCNN) for face detection in unconstrained conditions, while Li *et al.* [19] presented a method of integrating CNNs and a 3D face model in an end-to-end multi-task discriminative learning framework for face detection in the wild. Opitz *et al.* [23] proposed a novel grid loss layer for CNNs to deal with partial occlusion in face detection, which minimized the error rate on sub-blocks of a convolution layer independently rather than over the whole feature map. Chen *et al.* [4] proposed a new cascaded Convolutional Neural Network, named Supervised Transformer Network, to address the challenge of large pose variations in real-world face detection. In [25], Ranjan *et al.*presented Hyperface for simultaneous face detection, landmarks localization, pose estimation and gender recognition by using CNNs. It well exploits the synergy among the tasks which boosts up their individual performances.

To sum up, many datasets and models exist in the literatures, but few of them are especially developed for masked face detection. Toward this end, we introduce the dataset **MAFA** for training and benchmarking masked face detectors, while a new LLE-CNNs model is proposed as a baseline as well. Details of the proposed dataset and model will be described in the subsequent two sections.

3. MAFA: A Dataset of Masked Faces

In this section, we will first show the data collected for annotations in **MAFA**, and then describe the details of the annotation process. Finally, we will provide a brief statistic to show the major characteristics of **MAFA**.

3.1. Dataset Construction

We first collect a set of facial images from the Internet. In this process, keywords such as 'face, mask, occlusion and cover' are used to retrieve more than 300K images with faces from social networks like Flickr and the image search engines like Google and Bing. Note that we only keep the images with a minimal side length of 80 pixels. After that, images that contain only faces without occlusion are manually removed. Finally, we obtain $30,811$ images in total, and each image contains at least one masked face.

On these images, we ask nine subjects to manually annotate all faces, and each image is annotated by two subjects and cross-validated by the third subject. In the annotation process, we define six attributes that should be manually annotated for each face, including three types of locations and three types of face characteristics:

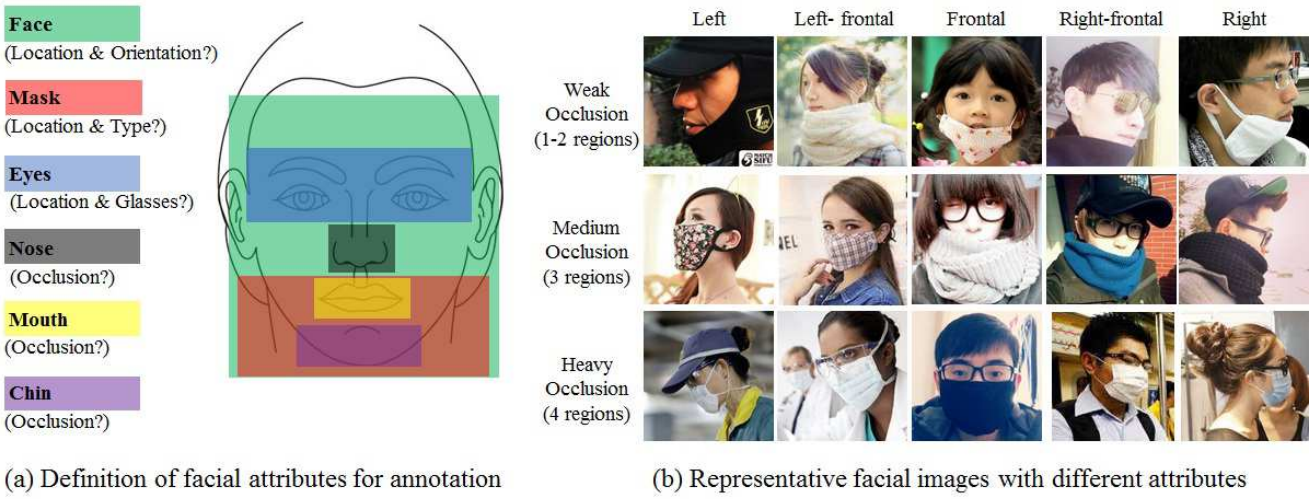
1) Location of faces. As in [13, 30], the location of each face is annotated by a square. Similar to [30], a face will be labeled as "Ignore" if it is very difficult to be detected due to blurring, severe deformation and unrecognizable eyes, or the side length of its bounding box is less than 32 pixels. Note that faces with the label "Ignore" will not be counted as true positives or false alarms once being detected.

2) Locations of eyes. For each face, the coordinates of eye centers are manually marked.

3) Locations of masks. The locations of all masks and/or glasses in a face are annotated with rectangles.

4) Face orientation. We define 5 orientations, including left, front, right, left-front and right-front. In the annotation, the orientation of a face is voted by three subjects.

5) Occlusion degree. To measure the occlusion degree, we divide a face into four major regions (see Fig. 2), including eyes, nose, mouth and chin. According to the number of regions occluded by masks and/or glasses, we define three occlusion degrees, including Weak Occlusion (one or two regions), Medium Occlusion (three regions), and Heavy Occlusion (four regions).



(a) Definition of facial attributes for annotation

(b) Representative facial images with different attributes

Figure 2. Definition of facial attributes and representative facial images in **MAFA**.

6) Mask type. We define four categories of masks that can be frequently found in Internet images, including: Simple Mask (man-made objects with pure color), Complex Mask (man-made objects with complex textures or logos), Human Body (face covered by hand, hair, etc.) and Hybrid Mask (combinations of at least two of the aforementioned mask types, or one of the aforementioned mask type with eyes occluded by glasses).

Figure 3 shows an example of the annotated face. We can see that our annotations actually consist of three locations (face, eyes and masks) and three face characteristics (orientation, occlusion degree and mask type). With these annotations, we can well describe the main characteristics of faces, masks as well as their correlations. Considering the 5 face orientations, 3 occlusion degrees and 4 mask types, the **MAFA** dataset covers 60 cases of masked faces in our daily scenarios. In this manner, **MAFA** can be used to provide a comprehensive benchmark of face detectors on all kinds of masked faces.

3.2. Dataset Statistics

After the annotation, we obtain $35,806$ masked faces with a minimum size of 32×32. The distribution of face sizes is shown in Fig. 4(a), and we find that a face in **MAFA** covers 143×143 pixels on average and spreads across a large variety of sizes. Figure 4(b)-(d) illustrate the statistics of face attributes, including face poses, occlusion degrees and mask types. We can see that most faces in **MAFA** are front faces, while few faces are purely left or right. In this manner, face detectors that rely on the accurate localization of eyes can be tested as well, while the challenging cases such as left-front and right-front faces can be used to further assess their robustness. In addition, most faces have medium occlusion, which are also the most common cases

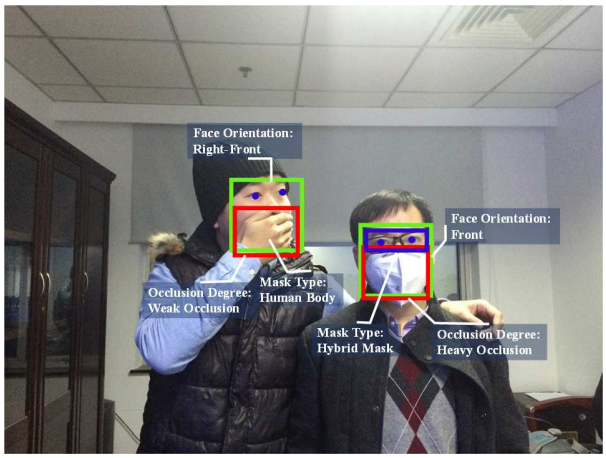


Figure 3. Examples of annotated faces in **MAFA**.

we may encounter in cold weather and poor air condition. Furthermore, **MAFA** contains multiple types of masks (*e.g.*, simple, complex and human body). These masks, which may bring in diversified 'noise' when constructing face descriptors, may provide us an opportunity to test the robustness of face detectors. To sum up, all these characteristics discussed above indicate that **MAFA** is a challenging dataset for face detection, and such a dataset can provide fair comparisons in benchmarking different face detectors

4. LLE-CNNs for Masked Face Detection

By inspecting the representative examples in **MAFA**, we find that two key challenges in masked face detection are the incomplete facial cues and inaccurate features from mask regions. For example, in a left-frontal face with the mouth, chin and nose occluded by a mask, most facial landmarks are invisible and the feature descriptor of the face will con-

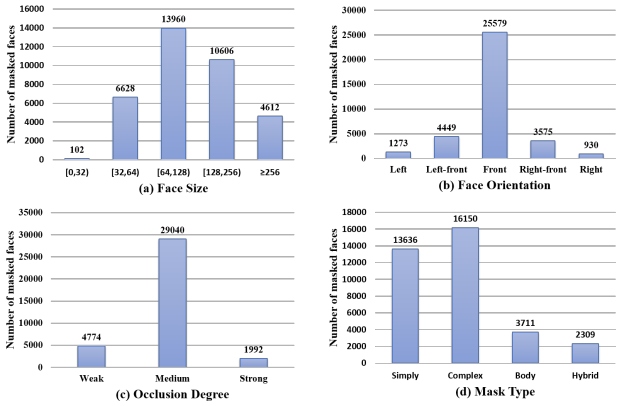


Figure 4. Statistics of the masked faces in **MAFA**.

tain a lot of noises generated from the mask regions. In this case, it is necessary to explore external cues beyond the face regions so as to recover the lost facial cues and alleviate the noisy mask features.

Toward this end, we propose KNN-CNNs for masked face detection, which consist of three major modules (as shown in Fig. 5), including: 1) the Proposal module extracts face proposals and describe them with noisy descriptors, 2) the KNN module refines such descriptor with respect to its nearest neighbors retrieved from a large pool of synthesized faces and non-faces, and 3) the Verification module jointly performs classification and regression tasks with a unified CNN to identify candidate facial regions and adjust their accurate positions. Details of the three modules are described as follows.

Proposal Module. The proposal module extracts and characterizes face candidates by cascading two CNNs for proposal generation and feature extraction, respectively. The proposal generation CNNs are built on the first four layers adopted in the P-Net of [35], which consists of three Convolutional layers and a softmax layer. Considering masked faces are very difficult to directly localize, we set a low threshold in the last layer to generate massive candidate proposals with such shallow CNNs. After that, we deliver each candidate region $\mathcal{F}_i$ into the pre-trained VGG-Face networks [24] to extract a 4096d descriptor from the `FC7` layer, which is then normalized to a vector $\mathbf{x}_i$. Since VGG-Face is trained on a large-scale face dataset, the extracted high-dimensional descriptor encodes both valuable facial cues as well as the noisy features from mask regions.

Embedding Module. The embedding module recovers the missing facial cues in $\mathbf{x}_i$ and suppresses the noisy features incorporated by mask regions. Toward this end, a feasible solution is to find the most similar faces or non-faces from an external database and use them to refine $\mathbf{x}_i$. Toward this end, we propose to embed a candidate region into a feature subspace formed by the most representative normal face and non-faces. Let $\mathbf{D}^+$ and $\mathbf{D}^-$ be two dictionaries consist of the descriptors of representative normal faces and non-faces, we can derive a new representation $\mathbf{v}_i$ from the noisy features of a candidate region through locally linear embedding,

$$\mathbf{v}_i = \arg\min_{\mathbf{v}} \|\mathbf{x}_i - \mathbf{D}_i\mathbf{v}\|_2^2, \quad \text{s.t. } \mathbf{v} \succeq 0. \tag{1}$$

where is $\mathbf{D}_i$ is a sub-dictionary that consists of the descriptors from the nearest neighbors of $\mathbf{x}_i$ in the columns of $\mathbf{D}_i$. Considering that it is somehow difficult to determine the optimal size of the sub-dictionary $\mathbf{D}_i$, we adopt a fast approximation of the LLE process by solving

$$\hat{\mathbf{v}}_i = \arg\min_{\mathbf{v}} \|\mathbf{x}_i - \mathbf{D}\mathbf{v}\|_2^2, \quad \text{s.t. } \mathbf{v} \succeq 0. \tag{2}$$

where $\mathbf{D} = [\mathbf{D}^+ \ \mathbf{D}^-]$ is a large dictionary that consists of both representative faces and non-faces, and the fixed dictionary $\mathbf{D}$ makes the quadratic optimization objects very easy to resolve. Thus the only question is, how to construct $\mathbf{D}^+$ and $\mathbf{D}^-$?

To construct the dictionaries $\mathbf{D}^+$ and $\mathbf{D}^-$, we refer to two previous datasets of normal faces [14, 32] and the training set of **MAFA** $25,876$ images. The same proposal module is applied to these images so as to obtain proposals, which are classified into four subsets, including normal faces $\mathbb{F}_N^+$ and non-faces $\mathbb{F}_N^-$ from images in [14, 32], and masked faces $\mathbb{F}_M^+$ and non-faces $\mathbb{F}_M^-$ from images the training set of **MAFA**. Considering that proposals can be very inaccurate due to the exitance of masks, we further extend each proposal in $\mathbb{F}_N^+$ and $\mathbb{F}_M^+$ by progressively crop rectangular windows that contain the top 25%, 50% and 75% area. In this manner, the sets of normal faces and masked faces can be extended to four times larger. After that, the descriptor of each face proposal is extracted by using the VGG-Face model as well.

Given these two sets, we assume that the columns of $\mathbf{D}^+$ and $\mathbf{D}^-$ consist of the descriptors from $\mathbb{F}_N^+$ and $\mathbb{F}_N^-$, respectively. In particular, $\mathbf{D}^+$ has the minimum error in sparsely encoding the masked faces in $\mathbb{F}_M^+$ and the maximum error in encoding the non-faces in $\mathbb{F}_M^-$, and vice versa for $\mathbf{D}^-$. In other words, $\mathbf{D}^+$ can be derived by solving

$$\begin{aligned} \min_{D^+} \ & \frac{1}{|\mathbb{F}_M^+|} \sum_{\mathcal{F}_1 \in \mathbb{F}_M^+} \|\mathbf{x}_1 - \mathbf{D}^+\alpha_1\|_2^2 \\ & - \frac{1}{|\mathbb{F}_M^-|} \sum_{\mathcal{F}_2 \in \mathbb{F}_M^-} \|\mathbf{x}_2 - \mathbf{D}^+\alpha_2\|_2^2 \\ \text{s.t. } & \|\alpha_1\|_0 = \|\alpha_2\|_0 = 1, \alpha_1 \succeq 0, \alpha_2 \succeq 0, \\ & \mathbf{D}_i^+ \in \{\mathbf{x}_0 | \mathcal{F}_0 \in \mathbb{F}_N^+\} \end{aligned} \tag{3}$$

where $\mathbf{x}_0$, $\mathbf{x}_1$ and $\mathbf{x}_2$ are the 4096d descriptors of $\mathcal{F}_0$, $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. $\mathbf{D}_i^+$ is the ith column of $\mathbf{D}^+$. α_1

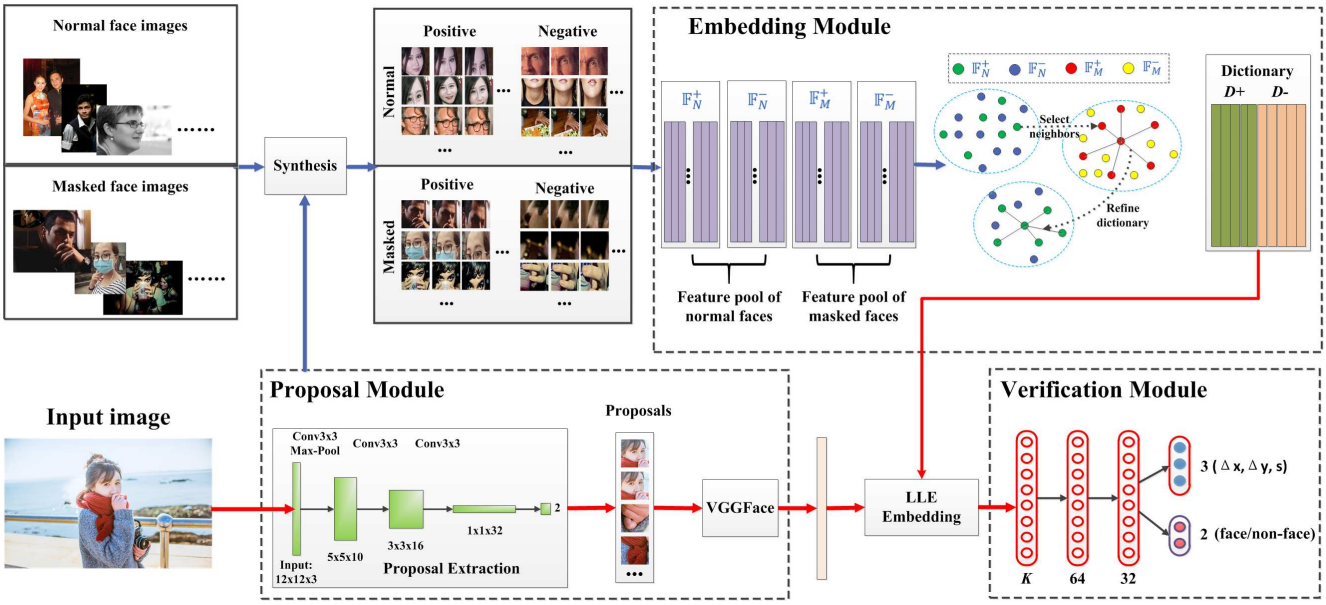


Figure 5. System framework of the proposed LLE-CNNs, which consists of three modules. The proposal module extracts face proposals and characterize them with noisy descriptors. After that, the embedding module transfers the noisy descriptors into similarity-based vectors by embedding the proposals into the feature subspace formed by representative normal faces and non-faces. Finally, the verification module jointly perform regression and classification tasks so as to refine the location and scale of a proposal and identify whether it is a real face.

and α_2 are sparse coefficient vectors in encoding $\mathbf{x}_1$ and $\mathbf{x}_2$ with $\mathbf{D}^+$. With the constraints $\|\alpha_1\|_0 = \|\alpha_2\|_0 = 1$ and $\alpha_1 \succeq 0, \alpha_2 \succeq 0$, the sparse coding process is equivalent to finding the nearest neighbor $\mathbf{D}^+$. Note that this optimization problem is different from the classic sparse coding scheme due to the constraint $\mathbf{D}_i^+ \in \{\mathbf{x}_0 | \mathcal{F}_0 \in \mathbb{F}_N^+\}$, making the problem difficult to resolve with classic optimization algorithms. Therefore, we propose a greedy algorithm to efficiently construct $\mathbf{D}^+$ from $\{\mathbf{x}_0 | \mathcal{F}_0 \in \mathbb{F}_N^+\}$.

In the proposed algorithm, we first compute the total cost of each proposal $\mathcal{F}_0 \in \mathbb{F}_N^+$ if it is solely used in encoding all masked faces and non-faces in $\mathbb{F}_M^+$ and $\mathbb{F}_M^-$:

$$\begin{aligned} \ell(\mathcal{F}_0) = & \min_{\{\rho_1\}} \sum_{\mathcal{F}_1 \in \mathbb{F}_M^+} \|\mathbf{x}_1 - \rho_1 \cdot \mathbf{x}_0\|_2^2 \\ & - \min_{\{\rho_2\}} \sum_{\mathcal{F}_2 \in \mathbb{F}_M^-} \|\mathbf{x}_2 - \rho_2 \cdot \mathbf{x}_0\|_2^2, \\ & \text{s.t. } \rho_1 \geq 0,\ \rho_2 \geq 0, \end{aligned} \tag{4}$$

By computing $\{\ell(\mathcal{F}_0) | \mathcal{F}_0 \in \mathbb{F}_N^+\}$, we obtain a ranked list of normal face proposals with ascending cost, in which the top-ranked proposals have the best capability in encoding masked faces and the worst capability in describing non-faces. As a result, we can construct $\mathbf{D}^+$ by iteratively move the top K proposals (we empirically set $K = 100$) from the ranked proposal list to a pool of selected proposals $\mathbb{P}$ whose descriptors are columns of $\mathbf{D}^+$. Let the initial $\mathbb{P}^{(0)} =$, we update $\mathbb{P}^{(t)}$ in the tth iteration by combining the newly selected top K proposals with $\mathbb{P}^{(t-1)}$. After that, the descriptors from proposals in $\mathbb{P}^{(t)}$ are used to update $\mathbf{D}^+$, which are then used to compute the objective score in (3). Note that the nth proposal in $\mathbb{P}^{(t)}$ will be removed if its descriptor is not used in the sparse encoding process. That is,

$$\sum_{\mathcal{F}_1 \in \mathbb{F}_M^+} \alpha_{1n} + \sum_{\mathcal{F}_2 \in \mathbb{F}_M^-} \alpha_{2n} = 0 \tag{5}$$

where α_{1n} and α_{2n} are the nth component of α_1 and α_2, respectively. By iteratively incorporating new proposals, encoding masked faces and non-faces and discarding unused ones, we can efficiently construct a dictionary that gradually minimize the lost in encoding masked faces and non-faces. The iteration will be stopped if such objective scores stop decreasing or reaches a predefined times T (we empirically set $T = 50$). After that, the dictionary $\mathbf{D}^+$ can be constructed with representative normal faces, and $\mathbf{D}^-$ can be obtained in a similar way.

Verification Module. The verification module classifies face candidates from the refined facial cues and refines their positions as well as scales. Given the dictionaries $\mathbf{D}^+$ and $\mathbf{D}^-$, the noisy descriptor of a candidate proposal can be converted into a similarity-based vector by using (2). After that, we deliver the vector into Deep Neural Networks that simultaneously perform the regression and classification tasks. The networks contain only fully convolutional layers, in which the regression branch output the refined location and scale of a proposal, and the classification branch identify the networks to be real face or not.

5. Experiments

In this Section, we conduct a comprehensive benchmark of our approach and 6 state-of-the-art face detectors on **MAFA**. The main objective of this benchmark is to find out how these face detectors perform in detecting masked faces and in which cases they will probably succeed and fail. By analyzing the experimental results of these face detectors, we then conduct a comprehensive discussion on how to develop face detectors with the capability of handling faces occluded by various kinds of masks.

5.1. Experimental Settings

As stated in Sect. 2, existing face detection models can be roughly grouped into three major categories (*i.e.*, boosting-based, DPM-based and CNN-based). From these three categories, we select six state-of-the-art face detectors for benchmarking, including two from the boosting-based category (**SURF** [18] and **NPD** [20]), three from the DPM-based category (**ZR** [37], **HH** [22] and **HPM** [7]), and one CNN-based detector (**MT** [35]).

In the benchmarking process, we split the **MAFA** into two subsets, including a training set and a testing set. The training set consists of $25,876$ images with $29,452$ masked faces that are randomly selected from **MAFA**, while the testing set contains the rest $4,935$ images with $6,354$ masked faces. In this study, we only report the performances of our approach and the six face detectors on the testing set, while such results can be directly used to facilitate their comparisons with new models trained on the same training set of **MAFA** in the future. The average precision (AP) scores of these face detectors are reported in Tab. 1, while some representative results are shown in Fig. 6.

5.2. Comparisons with State-of-the-Art Models

From Tab. 1, we can see that our approach significantly outperform the other six models in masked face detection. While our AP reaches up to 76.4% over the testing set of **MAFA**, the second best model, **MT**, only reaches an AP of 60.8%. Moreover, our approach outperforms the other six models in detecting masked faces under all the 12 conditions in face orientation, degree of occlusion and mask type. These results validates how effective our approach is and how difficult the task of masked face detection is. In particular, it further proves the necessity of constructing such a dataset with diversified masked faces, which can be not only used for model benchmarking but also used as an additional training source in developing new face detectors.

By analyzing the performance of all seven models in handling faces with different attributes, we find that the performances of all face detectors decrease sharply on faces with strong occlusions. Although our approach performs the best, its AP only reaches 22.5% when heavy occlusion occurs. Moreover, side faces will degrade the performance as well, especially the purely left or right faces. On such faces, the AP scores of our model only reaches 17.2% and 14.3%, respectively. Furthermore, even the very simple masks have strong impact to the face detectors. To sum up, all these results imply that **MAFA** is a challenging dataset even for the state-of-the-art face detectors.

Table 1. Average Precision (%) on the Testing Set of **MAFA**

Attributes	SURF [18]	NPD [20]	ZR [37]	HH [22]	HPM [7]	MT [35]	OUR	Min ↑
Left	0.02	1.01	5.02	7.91	1.29	6.89	**17.2**	9.29
Left-Fr.	2.17	4.37	29.3	28.5	26.6	31.9	**61.7**	29.8
Front	19.7	16.9	45.5	51.6	64.4	62.2	**79.6**	15.2
Right-Fr.	1.93	2.34	13.8	20.4	18.9	20.2	**54.5**	34.1
Right	0.02	0.23	1.34	5.43	0.93	1.94	**14.3**	8.87
Weak	18.1	5.87	37.1	47.7	58.5	56.2	**75.8**	17.3
Medium	12.7	17.0	13.9	46.4	34.8	45.6	**67.9**	22.3
Heavy	0.05	0.52	7.12	5.59	5.31	5.24	**22.5**	15.4
Simple	10.7	12.8	39.3	45.3	54.7	51.6	**74.3**	19.6
Complex	11.8	8.52	33.3	42.1	46.1	48.2	**71.6**	23.4
Body	12.3	4.12	21.4	34.7	23.4	30.4	**62.0**	27.3
Hybrid	0.17	0.63	7.64	7.58	6.00	6.48	**24.2**	16.6
All	16.1	19.6	41.6	50.9	60.0	60.8	**76.4**	15.6

To further investigate the performance of these models, we refer to the representative results shown in Fig. 6 and focus on the failures produced by the seven models. From these results, we find that the feature robustness might be a major cause of the impressive performance of our approach. By refining the noisy face descriptors extracted via VGG-Face within the embedding module, the inherent characteristics of faces, even occluded by masks, can be captured by referring to the most representative normal faces (and non-faces). In this manner, the reference faces (and non-faces) provide us valuable cues in recognizing a real face and refining its location and scale. By contrast, the boosting-based approaches often rely on the heuristic local features directly extracted from the masked face regions, which are less robust when heavy occlusion occurs. As a result, the performances of the two boosting-based face detectors are very low since the local features extracted from masks other than the facial regions are used for face detection.

Actually, feature robustness is with the highest importance in detecting masked faces. For example, **NPD** [20] uses a scale invariant features called normalized pixel difference and achieves an average precision of 19.6% since such simple features are often not reliable in masked facial regions. On the contrary, all the three DPM-based approaches, including **ZR**, **HH** and **HPM**, adopt the local HoG (Histogram of Oriented Gradients) descriptor [2]. As their DPM-based frameworks can well capture the robust topological information of facial regions from a global perspective, their performances are much better than **NPD** even

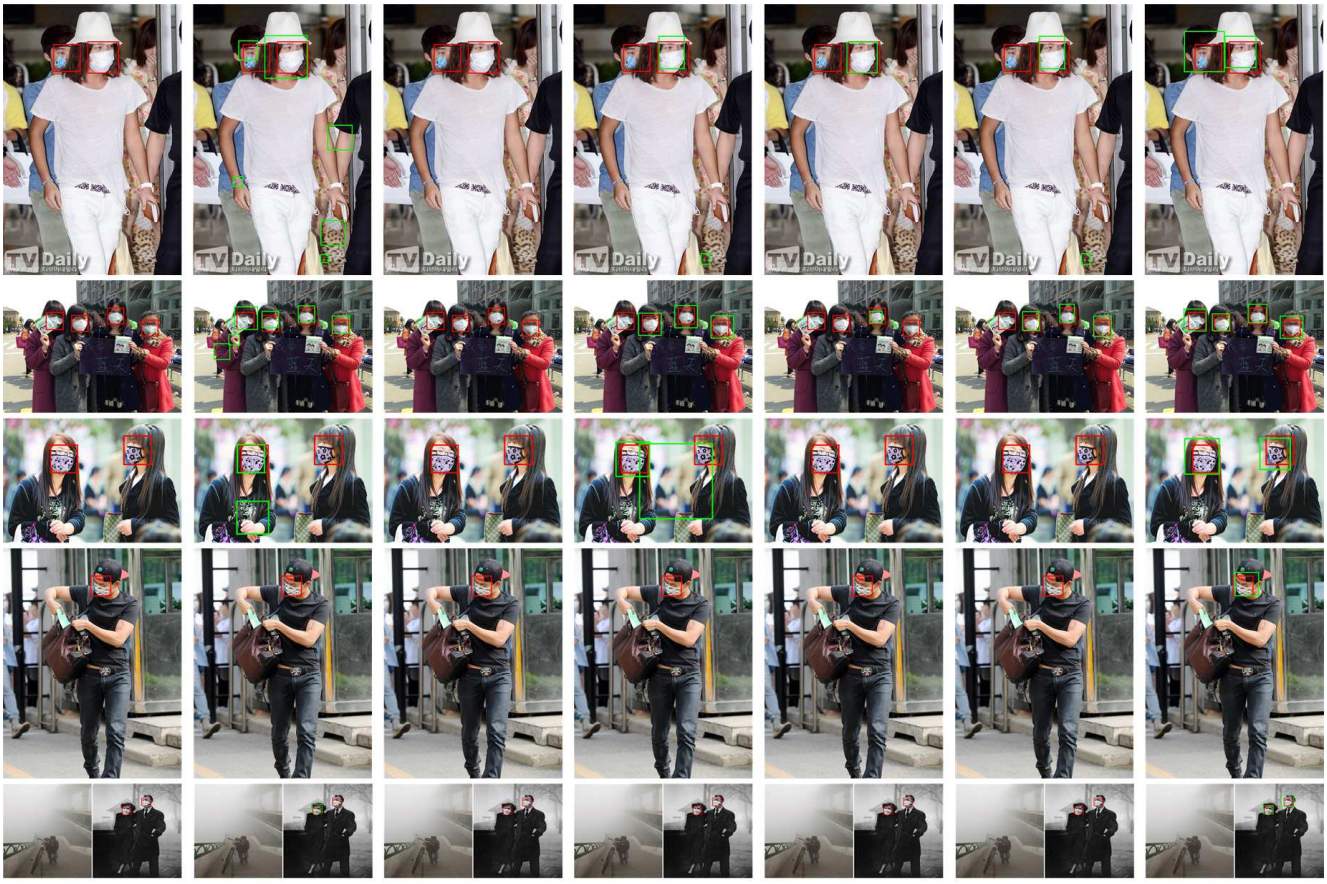


Figure 6. Representative detection results of our approach and 6 state-of-the-art face detectors on **MAFA** (Red: ground-truth, Green: detection results). We see that most face detectors have difficulties to accurately locate the side faces that are heavily occluded by masks.

in heavy occlusions and complex masks (see Fig. 6). Last but not least, the features in **MT** are learned from data. Even though such features are especially designed for masked faces, the overall performance of **MT** is still very impressive compared with the boosting-based and DPM-based approaches. To sum up, robust features play an important role in detecting masked faces, especially when most facial regions are occluded by masks. Compared with the heuristic features that are sensitive to masks and occlusions, robust topological feature or features learned from data are more robust in detecting faces. These results may point out a future direction of developing robust and effective face descriptors with the assistance of the **MAFA** dataset.

6. Conclusion

The problem of face detection in the wild have been explored in many existing researches, and the corresponding face detectors have been tested on datasets of normal faces. On these datasets, some face detectors have achieved extremely high performances and it seems to be somehow difficult to further improve them. However, the 'real wild' scenarios are much more challenging than expected for containing faces captured at unexpected resolution, illumination and occlusion. In particular, the detection of masked faces is an important task that needs to be addressed so as to facilitate applications such as video surveillance.

In this paper, we introduce the dataset **MAFA** and LLE-CNNs for masked face detection. We find that **MAFA** is very challenging for existing face detectors, while the proposed model achieves the best performance in all settings. This may imply that the data-driven framework may be a feasible solution in finding robust and effective features for masked face detection. We believe this dataset can facilitate the development of face detectors that can effectively detect faces with occlusions. In addition, predicting facial attributes of **MAFA** like mask type and occlusion degree also has practical meaning in many real-world applications, which will be one of our future research directions.

Acknowledgement. This work was partially supported by grants from National Key Research and Development Plan (2016YFC0801005), and National Natural Science Foundation of China (61672072 & 61402463).

References

[1] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *ECCV*, pages 109–122, 2014.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR*, pages 886–893, 2005.

[3] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE TPAMI*, 36(8):1532–1545, 2014.

[4] a. G. H. Dong Chen, F. Wen, and J. Sun. Supervised transformer network for efficient face detection. In *ECCV*, pages 122–138, 2016.

[5] S. S. Farfade, M. J. Saberian, and L. Li. Multi-view face detection using deep convolutional neural networks. In *ACM ICMR*, pages 643–650, 2015.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE TPAMI*, 32(9):1627–1645, 2010.

[7] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *IEEE CVPR*, pages 1899–1906, 2014.

[8] R. Girshick. Fast r-cnn. In *IEEE ICCV*, pages 1440–1448, 2015.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE CVPR*, pages 580–587, 2014.

[10] C. Huang, H. Ai, Y. Li, and S. Lao. High-performance rotation invariant multiview face detection. *IEEE TPAMI*, 29(4):671–686, 2007.

[11] V. Jain and E. L. Miller. FDDB: A benchmark for face detection in unconstrained settings. Technical report, University of Massachusetts, Amherst, 2010.

[12] B. Jun, I. Choi, and D. Kim. Local transform features and hybridization for accurate face and human detection. *IEEE TPAMI*, 35(6):1423–1436, 2013.

[13] B. Klare, B. Klein, E. Taborsky, and *et al.*. Pushing the frontiers of unconstrained face detection and recognition iarpa janus benchmark a. In *IEEE CVPR*, pages 1931–1939, 2015.

[14] M. Koestinger, P. Wohlhart, P. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[17] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *IEEE CVPR*, pages 5325–5334, 2015.

[18] J. Li and Y. Zhang. Learning SURF cascade for fast and accurate object detection. In *IEEE CVPR*, pages 3468–3475, 2013.

[19] Y. Li, B. Sun, T. Wu, and Y. Wang. Face detection with end-to-end integration of a convnet and a 3d model. In *ECCV*, pages 122–138, 2016.

[20] S. Liao, A. Jain, and S. Z. Li. A fast and accurate unconstrained face detector. *IEEE TPAMI*, pages 234–778, 2015.

[21] N. Markus, M. Frljak, I. S. Pandzic, J. Ahlberg, and R. Forchheimer. A method for object detection based on pixel intensity comparisons organized in decision trees. *CoRR 2014*, 2014.

[22] M. Mathias, R. Benenson, M. Pedersoli, and L. V. Gool. Face detection without bells and whistles. In *ECCV*, pages 720–735, 2014.

[23] M. Opitz, G. Waltner, G. Poier, H. Possegger, and H. Bischo. Grid loss: Detecting occluded faces. In *ECCV*, pages 386–402, 2016.

[24] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, pages 1–12, 2015.

[25] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *CoRR*, abs/1603.01249, 2016.

[26] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[27] I. Shlizerman, E. Shechtman, R. Garg, and S. M. Seitz. Exploring photo-bios. *ACM TOG*, 30:61, 2011.

[28] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE CVPR*, pages 511–518, 2001.

[29] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Convolutional channel features. In *IEEE ICCV*, pages 82–90, 2015.

[30] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Fine-grained evaluation on face detection in the wild. In *IEEE FG*, volume 1, pages 1–7, 2015.

[31] S. Yang, P. Luo, L. C. Change, and X. Tang. Wider face: A face detection benchmark. In *IEEE CVPR*, pages 5525–5533, 2016.

[32] S. Yang, P. Luo, C. C. Loy, and X. Tang. From facial parts responses to face detection: A deep learning approach. In *IEEE ICCV*, pages 3676–3684, 2015.

[33] S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild: past, present and future. *CVIU*, 138:1–24, 2015.

[34] C. Zhang and Z. Zhang. Improving multiview face detection with multi-task deep convolutional neural networks. In *IEEE WACV*, pages 1036–1041, 2014.

[35] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[36] C. Zhu, Y. Zheng, K. Luu, and M. Savvides. CMS-RCNN: contextual multi-scale region-based CNN for unconstrained face detection. *CoRR*, abs/1606.05413, 2016.

[37] X. Zhu and D. Ramanan. Face detection, pose estimation,and landmark localization in the wild. In *IEEE CVPR*, pages 2879–2886, 2012.