Face Detection, Alignment, Quality Assessment and Attribute Analysis with Multi-Task Hybrid Convolutional Neural Networks

GUO Da[1,2*], ZHENG Qingfang[3,4*], PENG Xiaojiang[1,2], and LIU Ming[3,4]

(1. Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. ZTE Corporation, Shenzhen, Guangdong 518057, China;
4. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen, Guangdong 518057, China)

Abstract: This paper proposes a universal framework, termed as Multi-Task Hybrid Convolutional Neural Network (MHCNN), for joint face detection, facial landmark detection, facial quality, and facial attribute analysis. MHCNN consists of a high-accuracy single stage detector (SSD) and an efficient tiny convolutional neural network (T-CNN) for joint face detection refinement, alignment and attribute analysis. Though the SSD face detectors achieve promising results, we find that applying a tiny CNN on detections further boosts the detected face scores and bounding boxes. By multi-task training, our T-CNN aims to provide five facial landmarks, facial quality scores, and facial attributes like wearing sunglasses and wearing masks. Since there is no public facial quality data and facial attribute data as we need, we contribute two datasets, namely FaceQ and FaceA, which are collected from the Internet. Experiments show that our MHCNN achieves face detection performance comparable to the state of the art in face detection data set and benchmark (FDDB), and gets reasonable results on AFLW, FaceQ and FaceA.

Keywords: face detection; face alignment; facial attribute; CNN; multi-task training

DOI: 10.12142/ZTECOM.201903004

http://kns.cnki.net/kcms/detail/34.1294.TN.20190920.2104.004.html, published online September 20, 2019

Manuscript received: 2019-06-11

1 Introduction

Face analysis has been widely-used in many applications such as face beautification system, face based access system, and video anti-terrorism system. Although great progresses have been made in recent, detecting and aligning abnormal faces such as occlusion faces as well as analyzing their attributes in surveillance are very challenging due to low resolution, lack of abnormal training data, etc.

Generally, face detection, face alignment, facial quality assessment, and facial attribute recognition are considered as separate face analysis tasks which may have their own task-dependent models. For face detection, from traditional Viola-Jones (VJ) face detector [1] and deformable part model (DPM) based face detector [2] to recent convolutional neural networks (CNN) based face detectors [3]–[9], the performance of face detection has been improved significantly. Among all the CNN based face detectors, those detectors evolved from anchor-based object detectors (e.g. single shot multibox detector (SSD) [10], Faster Region CNN (R-CNN) [11]), such as single shot scale-invariant face detector (S^3FD) [12] and Face Region CNN (R-CNN) [4], are superior to pure CNN face detectors [7], because anchor-based detectors can naturally leverage the context information. For face alignment, CNN based methods have also achieved promising results [12]–[16]. However, most of the alignment methods must be initialized by the provided face bounding box in advance, which presents a great demand of joint face and landmark detection [7], [17]. For facial quality assessment, traditional methods [18] mainly apply local binary patterns (LBP) [19] or histograms of oriented gradients (HOG)

This work was supported by ZTE Corporation and State Key Laboratory of Mobile Network and Mobile Multimedia Technology.
* Both authors contributed equally to this work.

[20] features with a support vector machine (SVM) classifier, while a few works with CNN obtain state-of-the-art performance [21], [22]. For facial attribute recognition, [23] introduces the CelebA dataset with 40 facial attributes ranging from smiling to gender, and subsequently many deep learning based methods are developed for facial attribute analysis [23]–[25]. Unfortunately, CelebA does not contain the attribute of wearing face mask which we are interested in.

In this paper, we address several face analysis tasks including face detection, face alignment, facial quality assessment, and facial attribute recognition in the wild. Specifically, we propose a Multi-Task Hybrid Convolutional Neural Network (MHCNN) which unifies all the tasks in a framework. MHCNN is comprised of two parts, namely a single stage detector (SSD) and an efficient tiny CNN (T-CNN). Compared to pure CNN face detectors, the SSD based face detector ensures high baseline accuracy on challenging face images in the wild. Instead of performing multi-task learning with SSD like [17], we apply a tiny CNN which is more feasible for multi-task face analysis. We argue that a CNN operated on cropped faces brings complementary information to SSD. Given an image, the MHCNN first detects all the faces with a SSD based face detector and then refines both the scores and bounding boxes with T-CNN. Since the T-CNN is applied in individual faces, it is straightforward to add multiple tasks upon it. We here add face alignment, facial quality assessment, and facial attribute recognition. In addition, we introduce a facial attribute dataset, i. e. FaceA, which contains two highly-concerned attributes in surveillance, namely wearing sunglasses and wearing masks. We also introduce a human-based facial quality assessment dataset, i. e. FaceQ, where low-quality cases include occlusion, low-resolution, large pose, etc. We evaluate our face detection performance on the well-known face detection data set and benchmark (FDDB) dataset, and demonstrate our T-CNN on our FaceA and FaceQ.

The remained of this paper is organized as follows. In Section 2, we review related work on face detection and multi-task learning. We introduce our MHCNN and its training strategy in Section 3. Our collected datasets are introduced in Section 4. We present experimental results in Section 5 and conclude the paper in Section 6.

2 Related work

We mainly review the face detection and multi-task learning for face analysis in this section. One can refer to [23]–[26] and [27] for face image quality assessment and facial attribute recognition, respectively.

2.1 Face Detection

Face detection has been a well-studied field of computer vision. According to the used features, face detection methods can be roughly divided into two categories, namely hand-craft feature based methods and CNN feature based methods.

1) Hand-craft feature based methods. The cascaded face detector proposed by VIOLA et al. [1] (VJ detector) obtains good performance in simple scenarios with real-time efficiency. Due to the relatively weakness of Haar-like features, the VJ detector degrades significantly in real-world applications with larger visual variations of human faces. Some works improved the VJ detectors by replacing the Haar-like features with more advanced hand-crafted ones [28]–[30], which need more computational cost. Another popular pipeline of face detection is based on DPM [2], [31], [32]. It performs relatively better than VJ detector in the wild but it is more computationally expensive and usually requires expensive annotation in the training stage.

2) CNN feature based methods. Since the remarkable success of CNN in object classification [33], many progresses have been made for face detection [3]–[9]. These CNN-based methods can be mainly concluded as three categories, namely cascaded CNN based face detection, two-stage region-based face detection, and single-stage face detection. The cascaded CNN based face detection pipeline, which inherits the advantage of the VJ detector, utilizes several small networks from simple to complex to detect faces and regress face boxes in a coarse-to-fine manner [5]–[7]. Two-stage region-based face detection pipeline is mainly transferred from region-based object detectors, like R-CNN [34], Fast R-CNN [35], and Faster R-CNN [11]. This method mainly includes two stages, namely proposal generation and classification. The single-stage face detection pipeline directly generates face boxes and scores from dense anchor boxes [8], [9]. The face detection model for finding tiny faces [7] trains separate detectors for different scales. S^3FD [12] presents multiple strategies to improve the performance of small faces. Single stage headless (SSH) [9] models the context information by large filters on each prediction module. PyramidBox [36] utilizes contextual information with improved SSD network structure. The advantage of single-stage face detectors is that it can use the context semantic information to assist in detecting faces, which is difficult for cascade face detectors. So, we introduce single-stage detection architecture into cascade face detector to get higher performance.

2.2 Multi-Task Learning

There are some existing works attempting to jointly solve the problem of face detection, alignment and facial attribute in a single model. YANG et al. [6] train deep convolution neural networks for facial attribute recognition to obtain high response in face regions which further yield candidate windows of faces. ZHANG et al. [37] proposed to use facial attribute recognition as an auxiliary task to enhance face alignment performance using deep convolutional neural network. CHEN [38] et al. apply random forest based on the features of pixel value difference to jointly conduct alignment and detection, but these handcraft features are low-level features and greatly limit its perfor-

mance. Multitask cascaded convolutional network (MTCNN) [7] leverages a cascaded architecture with three stages of shallow to deep convolution networks to jointly predict face and landmark locations in a coarse-to-fine manner, but the performance of MTCNN is limited by cascade architecture. So, we propose MHCNN with single-stage architecture in cascade face detector and joint multi-task learning to improve the performance of face detector.

3 Multi-Task Hybrid Convolutional Neural Networks

In this section, we first overview the proposed MHCNN, and then detail the two parts of MHCNN, namely the SSD-based face detector and the tiny multi-task CNN. We finally describe the training process of MHCNN.

3.1 Overview

Fig. 1 illustrates the pipeline of our MHCNN. The MHCNN consists of an SSD) and an efficient T-CNN. Given an image, the MHCNN first detects all the faces with the SSD based face detector and then refines both the scores and bounding boxes with T-CNN. T-CNN is also responded on facial landmark regression, facial quality assessment, and facial attribute recognition. We merge these facial tasks by considering that 1) it is more efficient than training multiple networks for each task, 2) some tasks, such as face classification and face attribute recognition, could be complementary with each other, and 3) T-CNN is performed on individual faces which could be complementary with an SSD-based face detector.

3.2 The SSD-Based Face Detector

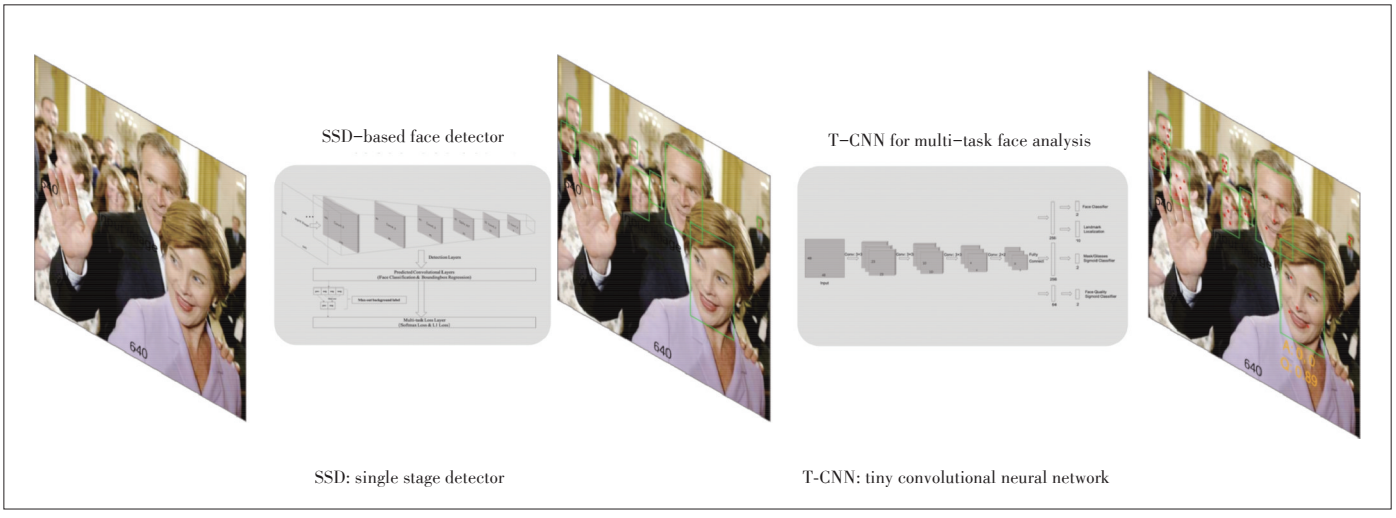
We resort to the recent S^{3}FD [12] as the first part of MHCNN since it is a state-of-the-art SSD-based face detector which is robust to face scale variation. **Fig. 2** shows the architecture of S^{3}FD. Both S^{3}FD and original SSD use VGG16 [39] as their backbone and pretrained on ImageNet. Compared to original SSD, S^{3}FD has several adjustments. We briefly review these adjustments as follows.

1) Anchor design. Instead of generating anchors with different scale and ratio for each detection layer in the original SSD, S^{3}FD generates one anchor scale for each detection layer since faces can be simply approximated as squares. Specifically, anchor scales range from 16 to 512 pixels with equal-proportion interval principle with strides increasing from 4 to 128 pixels. This strategy aims to guarantee that there are adequate features in different layers for detection.

2) Detection layers. One of the main challenges for face detection is to detect faces from tiny size to very large size. Unlike object detection of SSD in ImageNet, S^{3}FD moves forward the detection layers in order to detect tiny faces. Specifically, the detection layers of S^{3}FD are conv3_3, conv4_3, conv5_3, conv_fc7, conv6_2, and conv7_2. The norms of conv3_3, conv4_3, and conv5_3 are fixed to 10, 8, and 5 by L2 normalization for better training.

3) Max-out of background. Since negative anchors dominate the shallow layers with massive types, S^{3}FD adds a max-out layer for conv3_3 to relax the imbalance of positive and negative anchors. The max-out layer views the background label as several different labels and only takes the most active one for classification.

4) Scale compensation anchor matching. To match more tiny faces for anchors, S^{3}FD decreases the jaccard overlap threshold from 0.5 to 0.35 in order to increase the average number of matched anchors, and further sorts these anchors with jaccard overlap higher than 0.1 and selects top-N as matched anchors.



▲Figure 1. The pipeline of the proposed Multi-Task Hybrid Convolutional Neural Network (MHCNN). It consists of an SSD-based face detector for high-accuracy detection performance and a T-CNN for detection refinement and multi-task face analysis.

5) Training. We use the training set of the WIDER FACE [40] to train the SSD-based detector, and use the same data augmentation strategies as S^{3}FD, including color distort, random crop, and horizontal flip. The input size of network is fixed to 640×640. We use smooth L1 loss for face bounding box regression and softmax loss for face/non-face classification. We apply non-maximum suppression (NMS) to remove the highly overlapped results with a threshold of 0.7.

3.3 The Multi-Task Tiny CNN

We design an efficient T-CNN for the second part of MHCNN. The T-CNN aims to further refine the candidates from the SSD-based face detector, detect facial landmarks, assess the face quality, and recognize two importance facial attributes.

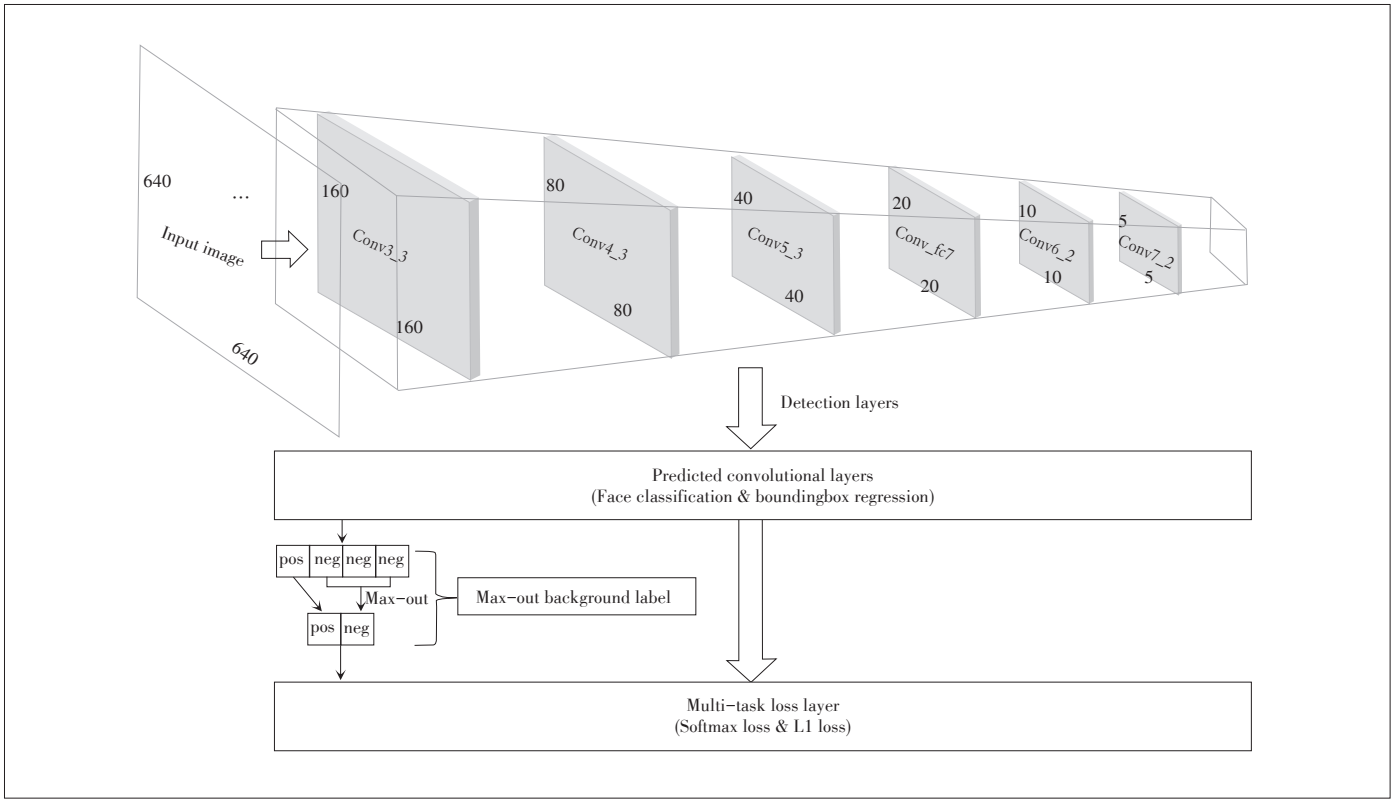
Fig. 3 presents the architecture of T-CNN. This architecture is inspired by the MTCNN [7]. Specifically, we use off-the-shelf O-Net of MTCNN as the architecture while add more tasks. T-CNN takes as input the 48×48 face regions, and output results for four face tasks as follows.

1) Face classification. We find that there are a number of non-face cases in the detections of the first part of MHCNN, which are mainly caused by hard negative contexts and low qualities, such as a person with back face. These cases can be relaxed by directly classifying the face regions. We believe that adding a refinement T-CNN could be complementary with the SSD-based face detector. The face scores from both S^{3}FD and T-CNN will be averaged to provide the final detections.

2) Landmark localization. We also predict five landmarks at eyes, nose, and mouth as in MTCNN. As shown in MTCNN, adding landmark localization is helpful for face recognition. We explain that landmarks can be viewed as a post validation of faces.

3) Attribute classification. Facial attribute is naturally a multi-label task. In this paper, we only concern about two important attributes for surveillance applications, i.e. wearing face masks and wearing sunglasses. A sigmoid layer is responded to each attribute.

4) Face quality classification. Face quality can impact the face/non-face classification scores in practice. We add face quality task as a two-class (i.e. high quality and low quality) classification problem since it is hard to annotate accurate quality scores for human. We consider two issues for face quality classification: a) face quality assessment served as a filter for subsequent face recognition system since too many low-quality and unrecognizable faces can impact the communication of front devices and cloud devices; b) As shown in Fig. 3, we use a separate fully-connected (FC) layer for face quality assessment because this task mainly depends on non-semantic information and it brings negative influence to other tasks if



▲**Figure 2. The architecture of single stage detector (SSD)-based face detector.**

▲**Figure 3. The architecture of tiny convolutional neural network (T-CNN).**

sharing the same FC layer in practice.

5) Training. Since T-CNN is partially served as a face/non-face refinement of the SSD-based detector in our MHCNN framework, we need to collect training data according to the results of SSD-based detector. To this end, we first calculate the Intersection-over-Union (IoU) ratio between the detections from the SSD-based detector and ground-truth faces on the training set of WIDER FACE, and then select these detections with IoU above 0.4 as positives and those less than 0.35 as negatives. The number of total face/non-face training data for T-CNN is 60000 with a ratio of 1:3. For facial landmark localization, we use the CelabA dataset which annotated with five facial landmarks, and apply random crop and gaussian blur as two data augmentation strategies. Euclidean loss is used for training facial landmark regression. We use our FaceA and FaceQ to train facial attribute recognition and face quality assessment. As for the training of facial attributes, the sigmoid layer with binary cross-entropy loss is used.

4 FaceA and FaceQ

This section details the collection and annotation of our FaceA and FaceQ datasets. To our knowledge, there is no public dataset for face attribute recognition with both wearing face masks and wearing sunglasses, and there is also no public dataset for face quality assessment in the wild. To meet our research, we collect the FaceA and the FaceQ datasets, and will make it publicly available to promote this area.

1) Collection. We make a python script and start from crawling "wearing sunglasses" and "wearing face mask" in image searching engine such as baidu (www.baidu.com) and google (www.google.com). We find it is hard to collect a large scale of data for wearing face mask since this case usually happens in surveillance. Totally, we crawled 1 409 and 1 335 images for "wearing sunglasses" and "wearing face mask", respectively. After crawling, we then feed these images into the first part of our MHCNN and crop the detected faces for further annotation.

2) FaceA. With the detected faces, we find there are many noises which are not human faces (e.g. cartoon and animation) or without the expected attributes. We manually remove these noises, and finally the FaceA dataset consists of 1 072 faces with sunglasses and 663 faces with face mask. The FaceA also includes a background category which contains 630 faces without wearing things. We randomly split both classes with 8:2 as training set and test set. **Fig. 4** shows some examples of FaceA. We note that these faces are mostly with large head poses which could be challenging for recognition.

3) FaceQ. With the crawled images, we find there are a lot of faces that neither belong to "wearing sunglasses" nor "wearing face mask", and that there are a number of faces with either high or low resolution. Thus, we collect FaceQ from the same source with FaceA but with three rules:

- Face resolution: Blurred and tiny faces are divided into low-quality class.

▲**Figure 4. Examples of our FaceA dataset.**

- Head pose: We collect faces as high-quality class only if their eyes can be seen clearly and their resolution are larger than 80×80. In other word, profile faces are not selected as high-quality class.
- Occlusion: Occluded faces are selected as the low-quality class except for the faces that are only occluded by wearing sunglasses and masks.

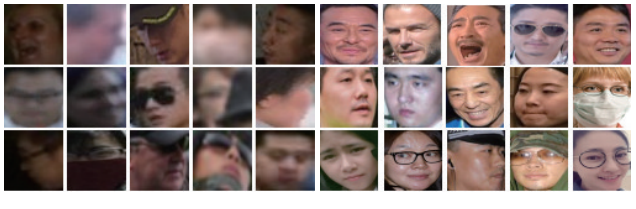
After manually selection, we totally obtain 1001 high-quality faces and 1 097 low-quality faces. We also randomly split both classes with 8:2 as training set and test set. Some examples of FaceQ are shown in **Fig. 5**.

5 Experiments

In this section we first present the implementation details, and then analyze the effectiveness of our joint multi-task training for the face detection, and further evaluate the final model on FDDB face detection benchmark and our own benchmarks. Finally, we evaluate the time of inference of our MHCNN.

5.1 Implementation Details

We use Caffe toolbox for implementation of our MHCNN. For the SSD-based face detector, we follow the training setting of S^{3}FD, using the pretrained VGG16 to initialize, and the other layers are randomly initialized with the "Xavier" method [13]. We fine-tune the pretrained model using stochastic gradient descent (SGD) with 0.9 momentum, 0.0005 weight decay and batch size 32. We train 80k iterations by using 10^{-3} learning rate, then continue training for 20k iterations with 10^{-4} and 10^{-5} learning rates. For T-CNN, we convert different datasets to hdf5 format for joint multi-task training, and train the model using SGD with 0.9 momentum and 0.0005 weight decay. The batch size of each task is 512, and we concatenate all data for joint training. Due to the fact that face quality assessment task depends on different information (i.e. low-level information) compared to the other tasks, we train T-CNN in two stages. First, we train the face classification, facial landmark localization, and face attribute recognition tasks with 10^{-1} learning rate for 200 iterations. Then we freeze the weights and only train the face quality recognition part of T-CNN with 10^{-1} learning rate for another 500 iterations.



a) Low-quality faces b) High-quality faces

▲Figure 5. Examples of our FaceQ dataset.

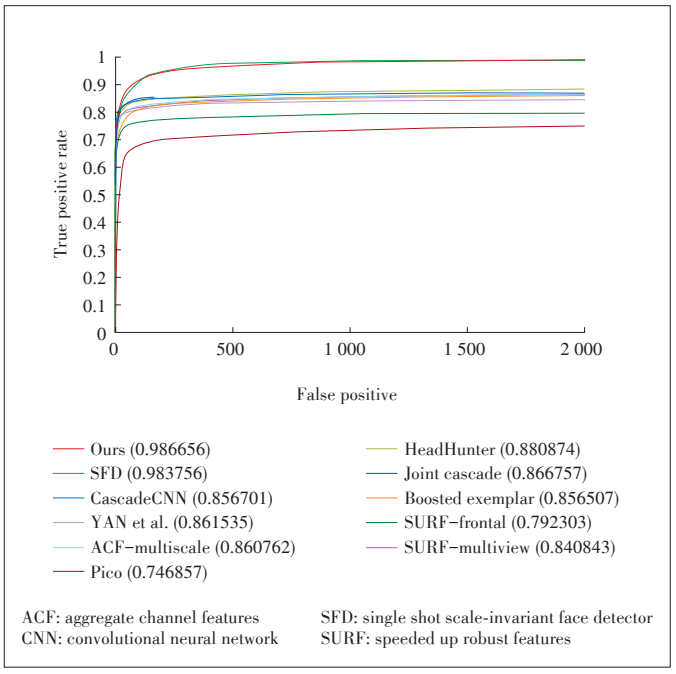
5.2 Evaluation on Face Detection Task

We evaluate and compare the face detection performance of our MHCNN on the FDDB dataset. FDDB contains 5 171 face annotations in 2 845 images. We compare our face detector to the state-of-the-art methods [2], [3], [12], [38], [41]–[45]. **Table 1** shows the recall ratio at 2 000 false positive and **Fig. 6** compares the receiver operating characteristic (ROC) curves to several state-of-the-art methods. Although our T-CNN aims for multi-task facial analysis, compared to the original S^{3}FD, our extra T-CNN provides complementary information for face/non-face classification which boosts S^{3}FD by around 0.3%. It is worth noting that a tiny improvement is difficult on the nearly-saturated FDDB. From Fig. 6, we observe that our MHCNN

▼Table 1. Comparison of our MHCNN on FDDB

Methods	Recall
Cascade CNN [3]	85.67%
ACF-multiscale [41]	86.08%
YAN et al. [42]	86.15%
Faster R-CNN [11]	96.10%
S^{3}FD [12]	98.37%
MHCNN	**98.66%**

ACF: Aggregate Channel Features
CNN: Convolutional Neural Network
MHCNN: Multi-task Hybrid CNN
R-CNN: Region CNN
S^{3}FD: Single Shot Scale-invariant Face Detector



ACF: aggregate channel features
CNN: convolutional neural network
SFD: single shot scale-invariant face detector
SURF: speeded up robust features

▲Figure 6. Discontinuous ROC curves on the FDDB dataset.

mainly improve the true positive rate at low false positive rate, which means the MHCNN has higher scores for true faces than S^3FD. This character is practical in real applications.

5.3 Ablation Study of T-CNN

We make an ablation study of our T-CNN on the facial attribute task. We perform experiments on the collected FaceA dataset, and use the sigmoid scores for each attribute with the best threshold searched on the training set.

Table 2 presents the results of ablation study on FaceA. We find several observations from Table 2. First, adding the facial landmark localization task improves both attribute tasks, where the gains for sunglasses and face mask are 0.43% and 2.6%, respectively. Second, training with all attribute tasks and landmark localization achieves the best results on FaceA. Third, the overall results with multiple tasks are relatively high though we only use a low-resolution input, which demonstrates the efficiency of our T-CNN.

1) Visualization on FaceA. **Fig. 7** visualizes some false positives on FaceA. We find that "wearing mask" is easily confused by large-pose faces with heavy hair, partial occlusion, and sunglasses; "wearing sunglasses" is confused by wearing common glasses, partial occlusion, and wearing mask.

2) Face quality assessment with T-CNN. For face quality assessment task, we evaluation our T-CNN on the FaceQ dataset. We compare a well-known and popular method in real applications, i.e. LBP feature with SVM. In this method we use the circular LBP operator with 8 sampling points in a circular area with radius 2, and divide face image into 7×5 to calculate LBP histogram, getting a 2 065 dim feature vector for each face image. Using the LBP face image features, we train a SVM model with radial basis kernel function (RBF) to predict either the normalized comparison scores. The cost of SVM is set at 1.5 and the gamma for RBF at 6.82. **Table 3** presents the comparison between MHCNN and LBP+SVM. We find that T-CNN outperforms LBP+SVM by 3.34%, which demonstrates its effectiveness. As a traditional method, LBP+SVM is also promising on this task which may be explained by that the face quality assessment task mainly depend on low-level texture information.

3) Visualization on FaceQ. **Fig. 8** shows some false positives on FaceQ. We find that most of the faces with serious occlusion by sunglasses and masks are recognized as low-quality faces, which may make sense since they are not suitable for recognition by human; several smooth profile faces also have low quality scores since they have little texture information; these low-quality faces with small occlusion by sunglasses are easily categorized into high-quality faces, which may be explained by the fact that these faces can provide relatively rich texture information.

5.4 Inference Time

During inference, we set 0.5 as the face confidence threshold in both parts of MHCNN. We perform the inference in 10 real-world surveillance images with 1 080×1 920 scales and report the average time. We first downscale the images to 320×568 and then use our MHCNN. The inference time of the SSD-

▼**Table 2. Ablation study of T-CNN on the FaceA dataset**

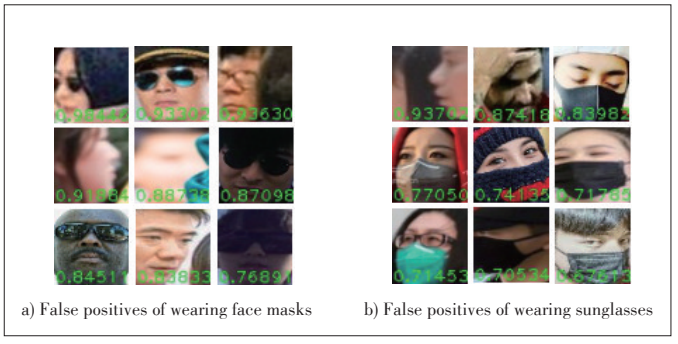
Methods (task setting)	Accuracy of Sunglasses (Threshold = 0.5)	Accuracy of Mask (Threshold = 0.5)
T-CNN (sunglasses)	76.14%	----
T-CNN (sunglasses + landmarks)	76.57%	----
T-CNN (masks)	----	83.30%
T-CNN (masks + landmarks)	----	85.90%
T-CNN (sunglasses + masks + landmarks)	**98.70%**	**99.35%**

T-CNN: tiny CNN

▼**Table 3. Evaluation on the FaceQ dataset**

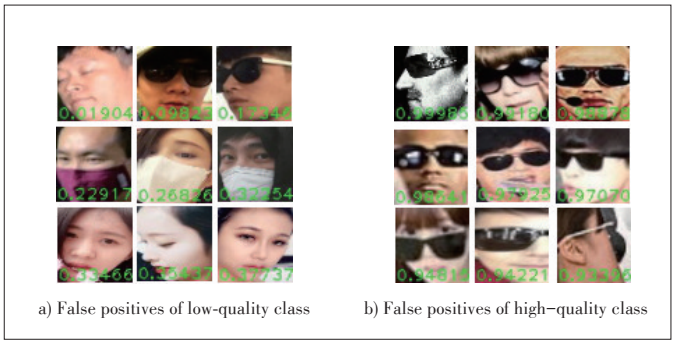
Methods	Accuracy of Face Quality (Best threshold)
LBP+SVM	78.52%
T-CNN	**81.86%**

LBP: local binary patterns SVM: support vector machine T-CNN: tiny CNN



a) False positives of wearing face masks b) False positives of wearing sunglasses

▲**Figure 7. False positives on FaceA test set. The score of images are the probability to predict wearing-mask and wearing-sunglasses.**



a) False positives of low-quality class b) False positives of high-quality class

▲**Figure 8. False positives on FaceQ test set. The faces with higher scores are predicted to the high-quality class and those with lower scores are predicted to the low-quality class.**

based detector and T-CNN are around 22 ms/frame and 2 ms/frame in NVIDIA TITAN Xp, respectively. Overall, our MHCNN can run 40 FPS in NVIDIA TITAN Xp for four face tasks including resizing computational time.

6 Conclusions

In this paper, we propose MHCNN for face detection, facial landmark detection, facial quality, and facial attribute analysis. We combine the single stage detector and CNN-based detector to boost the performance of face detection and implement multi-task learning. Our MHCNN achieves real-time performance in NVIDIA GPU for four face tasks. Additionally, we contribute two datasets on face attribute and face quality assessment. Experiments show that our MHCNN achieves the state-of-the-art on FDDB benchmark and gets reasonable results on FaceQ and FaceA.

References

[1] VIOLA P, JONES M J. Robust Real-Time Face Detection [J]. International Journal of Computer Vision, 2004, 57(2): 137–154. DOI: 10.1023 / B: VISI.0000013087.49260.fb
[2] MATHIAS M, BENENSON R, PEDERSOLI M, et al. Face Detection Without Bells and Whistles [C]//European Conference on Computer Vision. Zurich, Switzerland, 2014: 720–735. DOI: 10.1007/978-3-319-10593-2_47
[3] ZHU C, ZHENG Y, LUU K, et al. CMS-RCNN: Contextual Multi-Scale Region-Based CNN for Unconstrained Face Detection [M]. Deep Learning for Biometrics. Cham, Switzerland: Springer, 2017: 57–79
[4] JIANG H, LEARNED-MILLER E. Face Detection with the Faster R-CNN [C]// 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). Washington DC, USA, 2017: 650–657. DOI: 10.1109/FG.2017.82
[5] LI H, LIN Z, SHEN X, et al. A Convolutional Neural Network Cascade for Face Detection [C]//IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 5325–5334. DOI: 10.1109/CVPR.2015.7299170
[6] YANG S, LUO P, LOY C C, et al. From Facial Parts Responses to Face Detection: A Deep Learning Approach [C]//IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 3676–3684. DOI: 10.1109/ICCV.2015.419
[7] ZHANG K, ZHANG Z, LI Z, et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks [J]. IEEE Signal Processing Letters, 2016, 23(10): 1499–1503. DOI: 10.1109/LSP.2016.2603342
[8] ZHANG S, ZHU X, LEI Z, et al. Faceboxes: A CPU Real-Time Face Detector with High Accuracy [C]//2017 IEEE International Joint Conference on Biometrics (IJCB). Denver, Colorado, USA, 2017: 1–9. DOI: 10.1109 / BTAS.2017.8272675
[9] NAJIBI M, SAMANGOUEI P, Chellappa R, et al. SSH: Single Stage Headless Face Detector [C]//IEEE International Conference on Computer Vision. Venice, Italy, 2017: 4875–4884. DOI: 10.1109/ICCV.2017.522
[10] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot Multibox Detector [C]//European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 21–37. DOI: 10.1007/978-3-319-46448-0_2
[11] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [C]//Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 91–99. DOI: 10.1109 / TPAMI.2016.2577031
[12] ZHANG S, ZHU X, LEI Z, et al. S^3FD: Single Shot Scale-Invariant Face Detector [C]//IEEE International Conference on Computer Vision. Venice, Italy, 2017: 192–201. DOI: 10.1109/ICCV.2017.30
[13] GLOROT X, BENGIO Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks [C]//13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy, 2010: 249–256.
[14] SUN Y, WANG X, TANG X. Deep Convolutional Network Cascade for Facial Point Detection [C]//IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 3476–3483. DOI: 10.1109/CVPR.2013.446
[15] ZHU X, LEI Z, LIU X, et al. Face Alignment Across Large Poses: A 3D Solution [C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 146–155. DOI:10.1109/CVPR.2016.23
[16] FENG Z H, KITTLER J, AWAIS M, et al. Wing Loss for Robust Facial Landmark Localisation with Convolutional Neural Networks [C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 2235–2245. DOI: 10.1109/CVPR.2018.00238
[17] ZHUANG C, ZHANG S, LEI Z, et al. FLDet: A CPU Real-Time Joint Face and Landmark Detector [C]// IAPR International Conference on Biometrics (ICB). Crete, Greece, 2019
[18] BHARADWAJ S, VATSA M, SINGH R. Can Holistic Representations be Used for Face Biometric Quality Assessment? [C]//IEEE International Conference on Image Processing. Melbourne, Australia, 2013: 2792–2796. DOI: 10.1109 / ICIP.2013.6738575
[19] OJALA T, PIETIKÄINEN M, MÄENPÄÄ T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002 (7): 971–987. DOI: 10.1109/TPAMI.2002.1017623
[20] DALAL N, TRIGGS B. Histograms of Oriented Gradients for Human Detection [C]//International Conference on Computer Vision & Pattern Recognition (CVPR'05). San Diego, USA, 2005, 1: 886–893. DOI: 10.1109/CVPR.2005.177
[21] HERNANDEZ-ORTEGA J, GALBALLY J, FIERREZ J, et al. FaceQnet: Quality Assessment for Face Recognition Based on Deep Learning [DB/OL]. (2019-04-03). https://arxiv.org/abs/1904.01740
[22] NASROLLAHI K, MOESLUND T B. Face Quality Assessment System in Video Sequences [C]//European Workshop on Biometrics and Identity Management. Roskilde, Denmark, 2008: 10–18. DOI: 10.1007/978-3-540-89991-4_2
[23] LIU Z, LUO P, WANG X, et al. Deep Learning Face Attributes in the Wild [C]// IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 3730–3738. DOI: 10.1109/ICCV.2015.425
[24] HAN H, JAIN A K, WANG F, et al. Heterogeneous Face Attribute Estimation: A Deep Multi-Task Learning Approach [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(11): 2597–2609. DOI: 10.1109/TPAMI.2017.2738004
[25] RANJAN R, SANKARANARAYANAN S, CASTILLO C D, et al. An All-in-One Convolutional Neural Network for Face Analysis [C]//12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). Washington DC, USA, 2017: 17–24. DOI: 10.1109/FG.2017.137
[26] ZHANG Z, LUO P, LOY C C, et al. Learning Deep Representation for Face Alignment with Auxiliary Attributes [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(5): 918–930. DOI: 10.1109 / TPAMI.2015.2469286
[27] BEST-ROWDEN L, JAIN A K. Learning Face Image Quality from Human Assessments [J]. IEEE Transactions on Information Forensics and Security, 2018, 13(12): 3064–3077. DOI: 10.1109/TIFS.2018.2799585
[28] ZHANG L, CHU R, XIANG S, et al. Face Detection Based on Multi-Block LBP Representation [C]//International Conference on Biometrics. Seoul, South Korea, 2007: 11–18. DOI: 10.1007/978-3-540-74549-5_2
[29] ZHU Q, YEH M C, CHENG K T, et al. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients [C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR' 06). New York, USA, 2006, 2: 1491–1498. DOI: 10.1109/CVPR.2006.119
[30] PHAM M T, GAO Y, HOANG V D D, et al. Fast Polygonal Integration and its Application in Extending Haar-Like Features to Improve Object Detection [C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 942–949. DOI: 10.1109/CVPR.2010.5540117
[31] YAN J, LEI Z, WEN L, et al. The Fastest Deformable Part Model for Object Detection [C]//IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 2497–2504. DOI: 10.1109/CVPR.2014.320
[32] RAMANAN D, ZHU X. Face Detection, Pose Estimation, and Landmark Localization in the Wild [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Rhode Island, USA, 2012: 2879–2886. DOI: 10.1109 / cvpr.2012.6248014
[33] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet Classification with Deep Convolutional Neural Networks [C]//Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1097–1105. DOI: 10.1145/3065386
[34] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich Feature Hierarchies for

➡To P. 49

Frequency Range 300 MHz to 100 GHz [S], 2013.
[13] De Freitas P R, FILHO H T. Parameters Fitting to Standard Propagation Model (SPM) for Long Term Evolution (LTE) Using Nonlinear Regression Method [C]// IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA). Annecy, France, 2017: 84–88. DOI: 10.1109/CIVEMSA.2017.7995306
[14] RANI M S, BEHARA S, SURESH K. Comparison of Standard Propagation Model (SPM) and Stanford University Interim (SUI) Radio Propagation Models for Long Term Evolution (LTE), International Journal of Advanced and Innovative Research (IJAIR), 2012, 1(6): 221–228
[15] XU H, SHI C, ZHANG W, et al. Field Testing, Modeling and Comparison of Multi Frequency Band Propagation Characteristics for Cellular Networks [C]// IEEE International Conference on Communications. Kuala Lumpur, Malaysia, 2016. DOI: 10.1109/ICC.2016.7510961

Biographies

PEI Dengke (pdke1@mail.ustc.edu.cn) received the B.S. degree from China University of Geosciences, China in 2016. She is currently pursuing the B.E. degree at University of Science and Technology of China. Her research interest is localization and deep learning.

XU Xiaodong received his B.Eng. degree and Ph.D. degree in electronic and information engineering from University of Science and Technology of China (USTC) in 2000 and 2007, respectively. Since 2007, he has been a faculty member with the Department of Electronic Engineering and Information Science, USTC, where he is currently an associate professor. His research interests include array signal processing, wireless communications, and information-theoretic security.

QIN Xiaowei received the B.S. and Ph.D. degrees from the Department of Electrical Engineering and Information Science, University of Science and Technology of China (USTC) in 2000 and 2008, respectively. Since 2014, he has been a member of staff in Key Laboratory of Wireless-Optical Communications of Chinese Academy of Sciences at USTC. His research interests include optimization theory, service modeling in future heterogeneous networks, and big data in mobile communication networks.

LIU Dongliang received the B.S. degrees from the Telecommunications Engineering College, Beijing University of Posts and Telecommunications (BUPT), China, in 2004. Since 2006, he has been an Algorithms Researcher of the Algorithms Department in ZTE Corporation. His research interests include positioning theory, network optimization theory and big data in mobile communications.

ZHAO Chunhua received the M.S. degrees from the College of Electronic and Communication Engineering, Harbin Institute of Technology (HIT), China in 2002. Since then, she has been working on RRM algorithms design and network optimization in mobile communications. She has been an algorithms researcher at the Algorithms Department of ZTE Corporation since 2009. Her research interest is big data analytics for network optimizationeas of experimental regions are chosen as 100×100 in mobile communications.

←From P. 22

Accurate Object Detection and Semantic Segmentation [C]//IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 580–587. DOI: 10.1109/CVPR.2014.81
[35] GIRSHICK R. Fast R-CNN [C]//IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1440–1448. DOI: 10.1109/ICCV.2015.169
[36] TANG X, DU D K, HE Z, et al. Pyramidbox: A Context-Assisted Single Shot Face Detector [C]//European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 797–813. DOI: 10.1007/978-3-030-01240-3_49
[37] ZHANG Z, LUO P, LOY C C, et al. Facial Landmark Detection by Deep Multi-Task Learning [C]//European Conference on Computer Vision. Zurich, Switzerland, 2014: 94–108. DOI: 10.1007/978-3-319-10599-4_7
[38] CHEN D, REN S, WEI Y, et al. Joint Cascade Face Detection and Alignment [C]//European Conference on Computer Vision. Zurich, Switzerland, 2014: 109–122. DOI: 10.1007/978-3-319-10599-4_8
[39] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition [DB/OL]. (2014-09-04). https://arxiv.org/abs/1409.1556
[40] YANG S, LUO P, LOY C C, et al. Wider Face: A Face Detection Benchmark [C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vega, USA, 2016: 5525–5533. DOI: 10.1109/CVPR.2016.596
[41] YANG B, YAN J, LEI Z, et al. Aggregate Channel Features for Multi-View Face Detection [C]//IEEE International Joint Conference on Biometrics. Clearwater, USA, 2014: 1–8. DOI: 10.1109/BTAS.2014.6996284
[42] YAN J, ZHANG X, LEI Z, et al. Face Detection by Structural Models [J]. Image and Vision Computing, 2014, 32(10): 790–799. DOI: 10.1016/j.imavis.2013.12.004
[43] MARKUS N, FRLJAK M, PANDZIC I S, et al. Object Detection with Pixel Intensity Comparisons Organized in Decision Trees [DB/OL]. (2013-05-20). https://arxiv.org/abs/1305.4537
[44] LI H, LIN Z, BRANDT J, et al. Efficient Boosted Exemplar-Based Face Detection [C]//IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1843–1850. DOI: DOI: 10.1109/CVPR.2014.238
[45] LI J, ZHANG Y. Learning Surf Cascade for Fast and Accurate Object Detection [C]//IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 3468–3475. DOI: 10.1109/CVPR.2013.445

Biographies

GUO Da received the B.Eng. from the Computer Engineering College, JiMei University, China in 2018. He is currently a master student at the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China. His research direction is face detection and recognition based on deep learning.

ZHENG Qingfang received the B.S. degree in civil engineering from Shanghai Jiao Tong University, China in 2002 and Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Science, China in 2008. He is currently the chief scientist of video technology with ZTE Corporation. His research interests include computer vision, multimedia retrieval, image/video processing, with a special focus on low power embedded application and large-scale cloud application.

PENG Xiaojiang (xj.peng@siat.ac.cn) received his Ph.D. from School of Information Science and Technology from Southwest Jiaotong University, China in 2014. He currently is an associate professor at the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China. He was a postdoctoral researcher at Idiap Institute, Switzerland from 2016 to 2017, and was a postdoctoral researcher in LEAR Team, INRIA, France, working with Prof. Cordelia Schmid from 2015 to 2016. He serves as a reviewer for IJCV, TMM, TIP, CVPR, ICCV, AAAI, IJCAI, FG, Image and Vision Computing, IEEE Signal Processing Letter, Neurocomputing, etc. His research focus is in the areas of action recognition and detection, face recognition, facial emotion analysis, and deep learning.

LIU Ming received the M.Sc. degree from Harbin Engineering University, China in 2011. He is currently a senior engineer with ZTE Corporation. His research interests include object detection, tracking and recognition.