Task-Oriented Feature-Fused Network With Multivariate Dataset for Joint Face Analysis

Xuxin Lin, Jun Wan, *Member, IEEE*, Yiliang Xie, Shifeng Zhang, Chi Lin, Yanyan Liang, *Member, IEEE*, Guodong Guo, *Senior Member, IEEE*, and Stan Z. Li, *Fellow, IEEE*

***Abstract*—Deep multitask learning for face analysis has received increasing attentions. From literature, most existing methods focus on optimizing a main task by jointly learning several auxiliary tasks. It is challenging to consider the performance of each task in a multitask framework due to the following reasons: 1) different face tasks usually rely on different levels of semantic features; 2) each task has different learning convergence rate, which could affect the whole performance when joint training; and 3) multitask model needs rich label information for efficient training, but existing facial datasets provide limited annotations. To address these issues, we propose a task-oriented feature-fused network (TFN) for simultaneously solving face detection, landmark localization, and attribute analysis. In this network, a task-oriented feature-fused block is designed to learn task-specific feature combinations; then, an alternative multitask training scheme is presented to optimize each task with considering of their different learning capacities. We also present a large-scale face dataset called JFA in support of proposed method, which provides multivariate labels, including face bounding box, 68 facial landmarks, and 3 attribute labels (i.e., apparent age, gender, and ethnicity). The experimental results suggest that the TFN outperforms several multitask models on the JFA dataset. Furthermore, our approach achieves competitive performances on WIDER FACE and 300W dataset, and obtains state-of-the-art results for gender recognition on the MORPH II dataset.**

***Index Terms*—Attribute analysis, face analysis, face detection, landmark localization, multitask learning.**

I. Introduction

As a widely studied topic in computer vision, automatic face analysis usually consists of several different face tasks, such as face detection, facial landmark localization, and facial attribute analysis. In most real-world applications, these tasks are organized into a pipeline to execute step by step like facial expression recognition [1], [2]; age and gender recognition [3], [4]; and face verification [5]. These works rely on well-cropped face images as input or accurate facial landmarks used for face alignment. However, such a pipeline manner is not desirable since it brings the following drawbacks.

1) With the preprocessing of face detection and facial landmark localization, the performance of model heavily relies on off-the-shelf face detector and facial landmark detector.
2) These face tasks are usually heterogeneous but subtly correlated with each other. Since pipeline model optimizes different algorithms independently for each task, it cannot efficiently exploit the intrinsic correlation among these tasks.
3) For some algorithms with learned features like convolutional neural network (CNN), they cannot share features among different tasks in a pipeline framework, which will cause additional computing consumption.

In recent years, with the development of deep learning [6], [7] and multilabel learning [8] on image classification, deep multitask learning for face analysis has received increasing attentions [9]–[13]. In contrast to pipeline model, deep multitask method can avoid additional image processing when making an inference, and simultaneously resolve different face tasks with shared feature maps. However, most existing works mainly optimize a specific task in multitask manner, such as face detection [9], [14], [15]; facial landmark localization [10]; and facial attribute recognition [11]. Jointly optimizing all tasks and assuring the good performance of each task is challenging because of the following reasons.

1) Different face tasks usually require different levels of semantic features. For example, some fine-grained tasks

Manuscript received March 7, 2018; revised January 24, 2019; accepted May 7, 2019. Date of publication June 5, 2019; date of current version January 21, 2020. This work was supported in part by the National Key Research and Development Plan under Grant 2016YFC0801002, in part by the Chinese National Natural Science Foundation Projects under Grant 61876179 and Grant 61872367, and in part by the Science and Technology Development Fund of Macau under Grant 152/2017/A, Grant 0025/2018/A1, and Grant 008/2019/A1. This paper was recommended by Associate Editor J. Su. *(Xuxin Lin and Jun Wan contributed equally to this work.) (Corresponding author: Yanyan Liang.)*

X. Lin was with the Center for Biometrics and Security Research, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. He is now with the Faculty of Information Technology, Macau University of Science and Technology, Macau 999078, China (e-mail: linxuxin6@gmail.com).

J. Wan, S. Zhang, and S. Z. Li are with the Center for Biometrics and Security Research, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jun.wan@nlpr.ia.ac.cn; shifeng.zhang@nlpr.ia.ac.cn; szli@nlpr.ia.ac.cn).

Y. Xie and Y. Liang are with the Faculty of Information Technology, Macau University of Science and Technology, Macau 999078, China (e-mail: microos316@gmail.com; yyliang@must.edu.mo).

C. Lin is with the USC Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: linchi@usc.edu).

G. Guo is with the Institute of Deep Learning, Baidu Research, Beijing 100193, China, and also with the National Engineering Laboratory for Deep Learning Technology and Application, Baidu Research, Beijing 100193, China (e-mail: guoguodong01@baidu.com).

Digital Object Identifier 10.1109/TCYB.2019.2917049

like face bounding box regression and landmark regression tend to utilize low-level local features, which keep the facial position information. In contrast, high-level global features with class-specific information are suited to the coarse-grained tasks like face/nonface decision and facial attribute recognition. However, existing multitask models like [10]–[12] usually use a set of shared facial features from final fully connected layer before the decision of all tasks, which cannot fully consider the specific property of each task.

2) Owing to different learning capacities of each task, they have different convergence rate when joint learning. From the views of Zhang *et al.* [10], some tasks are possible to be overfitting earlier than the others during multitask training, which could affect the performance of the entire model. Although the work [10] introduces a task-wise early stopping strategy to facilitate learning convergence of model, the method need to predefine the importance coefficient of each task with partiality to main task.
3) Multitask training usually relies on rich label information. Most existing facial datasets provide limited labels for solving a specific task, such as PASCAL FACE [16], FDDB [17], and IJB-A [18] for face detection; LFPW [19], HELEN [20], and AFW [21] for facial landmark location; and CACD [22], IMDB-WIKI [23], and MORPH II [24] for facial attribute analysis. These datasets with restrictive labels would hinder the development of multitask model for joint face analysis.

To this end, we first introduce the JFA dataset, a large-scale face dataset with multivariate labels information, including face bounding boxes, 68 facial landmark, and 3 attribute labels: 1) apparent age (101 classes); 2) gender (2 classes); and 3) ethnicity (3 classes). The new dataset consists of 259 448 Internet images with 687 225 human faces. To the best of our knowledge, the JFA dataset is currently the largest face dataset that can be used to train multitask model for joint face analysis. To validate the performance of multitask model, we provide a set of complete evaluation protocols and metrics to assess each face task. The new dataset will be made publicly available and become an important complement of existing face datasets, which we believe would facilitate the development of multitask model for joint face analysis.

In the second part of this paper, we present a novel multitask network for joint face detection, landmark localization, and attribute analysis (i.e., age estimation, gender, and ethnicity recognition). Since the CNN is the powerful visual model that yields hierarchies of features with different information as demonstrated in [53], we can exploit the fact that low-layer feature maps respond to edges and corners and have better localization properties, while high-layer features are class-specific with high semantic information. In our network, we design a task-oriented feature-fused block, which can fuse feature maps from different convolutional layers. With multiple task-specific loss functions, the overall network can learn different feature combinations adapted to each particular task. Moreover, an alternative multitask training scheme is presented to consider different learning capacity of each task. Through this strategy, all tasks can be treated fairly and trained with reduced effect caused by different convergence rates. To validate the effectiveness of the proposed network, we study several recent multitask models based on CNN and reimplement them as comparison methods on the JFA dataset. The experimental results suggest that our method has better performance than other multitask models. Furthermore, our approach also achieves competitive performances on WIDER FACE [25] and 300W [26] dataset, and gets state-of-the-art results for age estimation and gender recognition on MORPH II [24] dataset simultaneously.

TABLE I
COMPARISONS BETWEEN THE JFA DATASET AND SEVERAL COMMON FACE DATASETS

Dataset	#Image	#Face	Bounding box	Age	Gender	Ethnicity	#Landmarks
PASCAL FACE [16]	0.85k	1.3k	✓	-	-	-	-
FDDB [17]	2.8k	5.1k	✓	-	-	-	-
MALF [27]	5.2k	11.9k	✓	-	✓	-	-
AFLW [28]	21.9k	25.9k	✓	-	-	-	21
IJB-A [18]	24.3k	49.7k	✓	-	-	-	-
WIEDR FACE [25]	32.2k	393k	✓	-	-	-	-
MORPH II [24]	55.1k	55.1k	-	✓	✓	✓	-
XM2VTS	2.3k	2.3k	-	-	-	-	68
LFPW	1.4k	1.4k	-	-	-	-	29
HELEN	2.3k	2.3k	-	-	-	-	194
CACD [22]	163k	163k	-	✓	-	-	-
IMDB [23]	523k	523k	✓	✓	✓	-	-
JFA (Ours)	**259k**	**687k**	✓	✓	✓	✓	**68**

The remainder of this paper is organized as follows. Section II provides an overview of related works about existing face datasets and multitask methods for face analysis. Section III introduces the JFA dataset, including the procedure of data construction and data statistic. Section IV describes the detail of proposed task-oriented feature-fused network (TFN) and training scheme. Section V shows the experimental results on different datasets as well as the detailed analysis. Finally, we conclude this paper and give a brief summary in Section VII.

II. RELATED WORK

A. Face Datasets

In the literature, there have been quite a few face datasets proposed for solving a specific face task. As shown in Table I, the PASCAL FACE [16], FDDB [17], and IJB-A [18] datasets are usually used for face detection since they only provide the annotation of face bounding box. With well-cropped face images, the XM2VTS [29], LFPW [19], and HELEN [20] datasets provide different number of facial landmark points that can be used to train landmark localization model. However, since the accuracy of these original annotations is low, 300W [26] reannotates them under a unified protocol with 68 landmark points. For facial attributes analysis, it is usually categorized as the recognition of local attributes (e.g., eyeglass, mustache, big nose, etc.) and global attributes (e.g., age, gender, ethnicity, etc.) in terms of the facial region

in which they can be found. In this paper, we mainly study global attributes recognition as they are more valuable and challenging than local attributes. The CACD [22] and IMDB-WiKi [23] datasets are derived from IMDB or Wikipedia with limited attribute labels in which the former only provides age label, while the latter contains both age and gender annotations. The MORPH II [24] dataset collects a large quantity of face images with three attribute labels, including age, gender, and ethnicity. However, these images are captured in a constrained environment with limited availability in real-world applications. Recently, some works tend to construct an informative dataset with more annotations than early datasets, such as AFLW [28], WIDER FACE [25], and MALF [27], which provide multiple facial labels like positions, landmarks, gender, and poses. Compared with early datasets, recent datasets collect more images with various annotated faces in real-world scenarios. For example, in FDDB [17], there are only 2800 images with 5100 annotated faces. In contrast, AFLW [28] provides 21 997 images with 25 993 faces and is about eight times higher than FDDB [17] in the number of images, and WIDER FACE [25] collects 32 203 images, including 393 000 faces, which are 77 times larger than that in FDDB [17]. With the development of deep learning, the presence of large-scale datasets can enable the ability of complex networks like [6], which achieve extreme performance by learning discriminative representations directly from data.

B. Multitask Models for Face Analysis

In recent years, there has been increasing interest in the multitask method for face analysis. According to the motivation for the adoption of multitask manner, we can group existing multitask models roughly into two categories, called auxiliary multitask model and joint multitask model.

1) Auxiliary Multitask Model: In this category, the multitask model is usually used to optimize a main task by learning several auxiliary tasks, the performances of which are limited and not chiefly concerned. As one of the earliest auxiliary multitask models for face analysis, Zhang and Zhang [9] presented a multitask framework based on CNN to facilitate the performance of multiview face detection by jointly learning face/nonface decision, pose estimation, and facial landmark localization. This method exploits the fact that CNN can effectively learn associated features by simultaneously training different face tasks. Similarly, Huang *et al.* [14] and Li *et al.* [15] improved the accuracy of face detection by incorporating the facial landmarks location and the prediction of heat map of facial key-points, respectively, in multitask network. Another typical auxiliary multitask model was proposed by Zhang *et al.* [10], which learns pose estimation and three facial attributes recognition (i.e., smiling, wearing glasses, and gender) to improve landmark localization using shared features from the last layer. This paper studied the effect for the performance of main task when jointly training multiple tasks and presented a task-wise early stopping strategy to facilitate learning convergence of model. These works can efficiently exploit the correlation between main task and auxiliary tasks, and further promote the performance of main task by multitask learning. However, they usually need to carefully define the weight loss of each task in a multitask loss function with the preference to main task.

2) Joint Multitask Model: Different from the auxiliary multitask model, joint multitask model tends to consider the performance of each task in a multitask framework. One of the earliest joint multitask models for face analysis was proposed by Zhu and Ramanan [21], which jointly resolves face detection, pose estimation, and facial landmark localization by using a mixtures of trees with a shared pool of parts. In recent years, with the development of deep learning, joint multitask model based on CNN has gained increasing attentions. Zhang *et al.* [13] proposed a multitask cascaded CNN for joint face detection and alignment. Due to the adoption of cascaded structure, the model has to be divided into several independent parts for training, which is similar to pipeline model. Instead of cascaded structure, Ranjan *et al.* [12] presented an end-to-end multitask learning framework to simultaneously optimize face/nonface decision, landmark localization, pose estimation, and gender recognition. However, since the model relies on candidate face images as input, a region proposal generation algorithm (i.e., selective search [30]) is used beyond the multitask framework and cannot be jointly optimized with other tasks. More recently, region-based multitask models have been widely studied to unify the detection and recognition tasks, which are applied into various fields, such as object instance segmentation [31], joint vehicle analysis [32], and person reidentification [33]. Especially, similar to this paper, He *et al.* [34] presented a region-based multitask network for joint face detection and facial attribute analysis. However, this method does not fully exploit the hierarchical features in CNN but only use the shared feature maps from top convolutional layer for the decision of each task.

III. JFA Dataset

A. Data Construction

1) Construction Methodology: Images in the JFA dataset are collected from the IMDB[1] website, where we follow a list of names of actors and retrieve the movie posters and screenshots from their introduction pages. For efficient image annotation with high accuracy, we design a coarse-to-fine annotation scheme in the following four steps.

1) By using Face++ API, one of free APIs for developers, we can get coarse labels of all images, including face bounding boxes, 83 facial landmarks, and 3 facial attribute labels (age, gender, and ethnicity). In this step, there are some faces undetected due to various poses, heavy occlusion, small size, etc.
2) To reduce missing faces, we redetect all images by using a recently proposed face detector [35]. In this process, if the intersection-over-union (IoU) overlap of a new detected face is higher than 0.7 with any annotated faces, it will be saved and replaced the former one.
3) We follow the annotation rule of facial landmarks in 300W [26] and transform 83 landmarks into

[1] http://www.imdb.com/

Fig. 1. Samples of the JFA dataset. It contains all kinds of faces, for example, the faces of celebrity, reporter, and passerby with different genres, scales, occlusions, and illumination. Most of the recognizable faces are annotated with the face bounding box, landmarks, and three attributes (i.e., apparent age, gender, and ethnicity), while the heavily occluded and blurred faces are only labeled with the red and blue boxes, respectively.

68 landmarks by utilizing interpolation and landmark matching [36].

4) For getting accurate labels in the end, we refine all annotations and fill up missing labels manually in terms of an annotation policy detailed below.

2) Annotation Tool: After getting coarse labels of all images, we can do annotation efficiently by exploiting an online annotation tool.[2] The tool can distribute labeling tasks to annotators easily by Web and track the completion progress of every one behind the scenes. For labeling conveniently, all detected faces are cropped by a suitable zoom and displayed with some annotation options. Since the facial landmarks and attributes of detected faces with small size cannot be recognized, the tool will only retain the facial image with its minimal side length is larger than 30 pixels at least. For ensuring the quality of annotation, the tool can record the number of check for annotated faces and evenly assign them into each annotator.

3) Annotation Policy: To keep the uniformity of annotation, we make some rules for each type of label. For recognizable faces in the JFA dataset, the forehead, chin, and cheek can be contained in a bounding box tightly as shown in Fig. 1. If some faces are very difficult to recognize their attributes and landmarks due to the heavy occlusion and blur, they will be ignored with a flag but still keep the bounding box. For some profile faces with the partial missing, the landmarks are not been saved, but the attributes are kept when the face is clear enough. Since the real age of each face is hard to obtain, we estimate the apparent age by referring to the work [37], which adopts the mean of ages annotated by different annotators as the label. During annotation, all facial images are evenly assigned to 15 independent annotators, and cross-checked twice to avoid the human error as long as possible.

[2]The tool will be released together with JFA.

TABLE II
JFA DATASET CONSISTS OF TRAINING SET, VALIDATION SET, AND TESTING SET WITH DIFFERENT RATIOS OF BOUNDING BOXES, LANDMARKS, AND ATTRIBUTES (I.E., APPARENT AGE, GENDER, AND ETHNICITY)

JFA	#Images	#Faces	Bounding boxes	Landmarks	Age/ Gender/ Ethnicity
Training	239,640	635,992	100%	62.62%	65.37%
Validation	9,904	25,360	100%	61.00%	66.79%
Testing	9,904	25,873	100%	60.65%	66.09%
Total	**25,9448**	**68,7225**	**100%**	**62.48%**	**65.45%**

B. Data Statistic

1) Overview: To the best of our knowledge, the JFA dataset is currently the largest face dataset that can be used for joint face analysis. Comparing with several popular face datasets in Table I, the JFA dataset has the largest number of annotated faces with multivariate label information. From Fig. 1, we can see that the JFA dataset contains various faces with different genres, scales, occlusions, and illumination, for example, the first column of the figure shows three genres of faces (i.e., color photograph, poster, black, and white shots). As shown in Table II, the JFA dataset consists of 259 448 images with 687 225 annotated faces, and is divided into three subsets for training, validation, and testing, respectively. Due to various complex factors, such as heavy occlusion, low-resolution, and small size (i.e., 30 pixels and less), some faces cannot be recognized for full labeling. To ensure the same distribution for all labels in each subset, we randomly select images from JFA and keep the same proportion for each type of label. In the following, we will detail the property and distribution of every specific label.

2) Bounding Box: The bounding box tightly covers the forehead, chin, and cheek of a face to describe the face size (see Fig. 1). As shown in Fig. 2(a), the face size in the JFA dataset has a large range of variations from around 16 pixels to more than 512 pixels. We can find that most of faces with full labels have a large enough size due to more clear facial features than that of small faces. For some difficultly recognizable faces, the bounding box is still saved with a reasonable estimation, which can be used to further assess the robustness of face detection.

3) Other Annotations: Following the 300W [26], we use the same annotation protocol to annotate a face with 68 facial landmarks. From Table II, we can see that the number of faces with landmarks label is about 429 400 over 60% in the JFA dataset. In addition, we provide three global attributes (i.e., apparent age, gender, and ethnicity), which can be inferred in a whole face. The distribution of each attribute is shown in Fig. 2(b)–(d). We can see that the ages vary from 0 to 100 years, and are mainly distributed in the range from 25 to 55 years. Moreover, the distribution of gender is relatively balanced while that of ethnicity is extremely unbalanced due

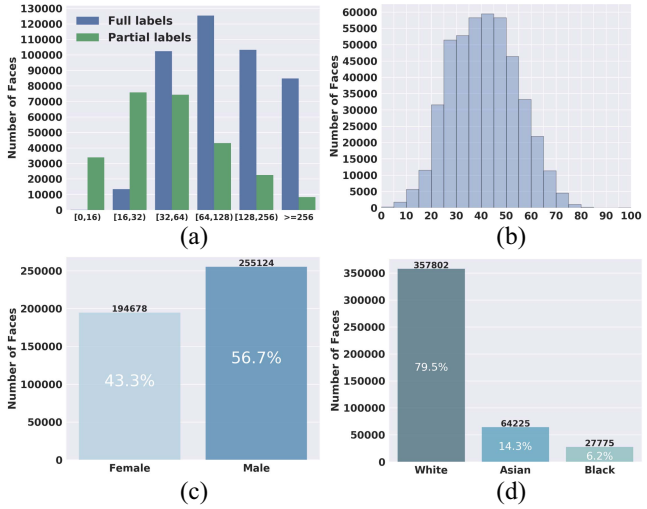


Fig. 2. Distribution of face size, age, gender, and ethnicity in the JFA dataset.

to the nature of data source. Since the facial attributes are recognized more easily than landmarks that heavily rely on visible regions of a face, we can infer the attributes annotation from a face with slight occlusion. From Table II, we can see that there are more annotated faces with attribute information than those with landmarks.

C. Evaluation Protocol and Metrics

1) Face Detection: By referring to the evaluation for detected bounding boxes in PASCAL VOC [38] and WIDER FACE [25], we provide the precision–recall curves with average precision (AP) to evaluate all detected faces. Note that if the detected face with the highest score has an IoU of more than 0.5 with any ground faces, it will be regard as a true prediction and false result otherwise.

2) Landmark Localization: Since the annotation of landmarks follows the rule of 300W [26], we also employ its evaluation protocol to assess the accuracy of predicted facial landmarks. We provide the cumulative error curve with the area under the curve (AUC), which describes the point-to-point root mean square (RMS) error normalized with the interocular distance. Moreover, we compute the average RMS error of every landmark (Avg.RMS) as a total index to evaluate the performance.

3) Attribute Analysis: For the evaluation of age estimation, gender, and ethnicity recognition, we refer to some related works [39]–[41] on MORPH II [24], which use the mean absolute error (MAE) to evaluate the accuracy of estimated age, while the accuracy rate is applied into the assessment of predicted gender and ethnicity.

IV. Task-Oriented Feature-Fused Network

A. Model Structure

1) Overall Framework: As shown in Fig. 3(a), we adopt the VGG-16 [7] as the base architecture of TFN, which contains 13 convolutional layers (i.e., conv1_1 to conv5_3 layers) and two fully connected layers (i.e., fc6 and fc7 layers). These convolutional layers are divided into five convolutional stacks in terms of their down-sampling ratios, and each of them in the same stack can generate feature maps with the same size. These convolutional stacks are exploited as the stem part of network and provide different levels of shared feature maps for each task. On the top of specific stacks, we design a multilayer assistant region proposal subnetwork (RPN) to generate various face region proposals. Following the multilayer RPN, three independent region-wise subnetworks are established for solving different face tasks. With the similar structure, each subnetwork consists of a task-oriented feature-fused block, two fully connected layers, and a decision layer with task-specific output. From Fig. 3(a), we can see the face detection net receives the face region proposals from the RPN and extract their feature maps from the feature-fused block. Through the decision layer, most of the invalid region proposals can be filtered, while other face proposals are refined with accurate location. For facial landmark localization and attribute analysis, instead of the proposals from RPN, new face boxes from face detection net are input into the following subnetworks for the landmark prediction and attribute recognition.

2) Multilayer Assistant RPN: Compared to the Faster R-CNN [42], in which an original RPN is established on the top convolutional layer of the network to generate region proposals responding to detected objects. In the proposed TFN, we design a multilayer assistant RPN on the top layers of the fourth and fifth convolutional stacks (i.e., conv4_3 and conv5_3 layers) to jointly catch various faces with different scales. In the RPN, with 256 channels and 3×3 kernel size, two convolutional layers are grafted on the top layers of the corresponding stacks, respectively, and generate specific feature maps, in which each location can be learned for face/nonface decision. As shown in Fig. 3(b), since the feature maps from high convolutional layer have smaller size but wider receptive field than those from low convolutional layer, we assign the reference boxes with the scales {128, 256, 512} into conv5_3 layer, while the conv4_3 layer is responsible for the reference boxes with the scales {16, 32, 64}. In addition, to consider the faces with different poses, each reference box has three aspect ratios {0.5, 1, 2}.

3) Task-Oriented Feature-Fused Block: In the TFN, instead of extracting single feature maps from the top convolutional layer, we design a customizable feature-fused block to fuse shared feature maps from different convolutional layers. As shown in Fig. 3(c), according to the position of proposals generated from RPN or face detection net, these shared global feature maps are cropped into facial feature maps in terms of the relative scaling between the input image and the global feature maps. Through region-of-interest (RoI) pooling defined in [43], these facial feature maps are resized with the same size of width and height. Since the feature maps from different layers have different scales of values, we normalize these values by using L2 normalization [44]. After normalizing, these feature maps are concatenated into the new feature maps with the sum of all channels. Finally, to unify the number of channels adapted to the following subnetworks, the dimension of feature maps is reduced by a 1×1 convolution and kept the same number of channels as that of the top convolutional layer. In the face detection net and attribute analysis net, the

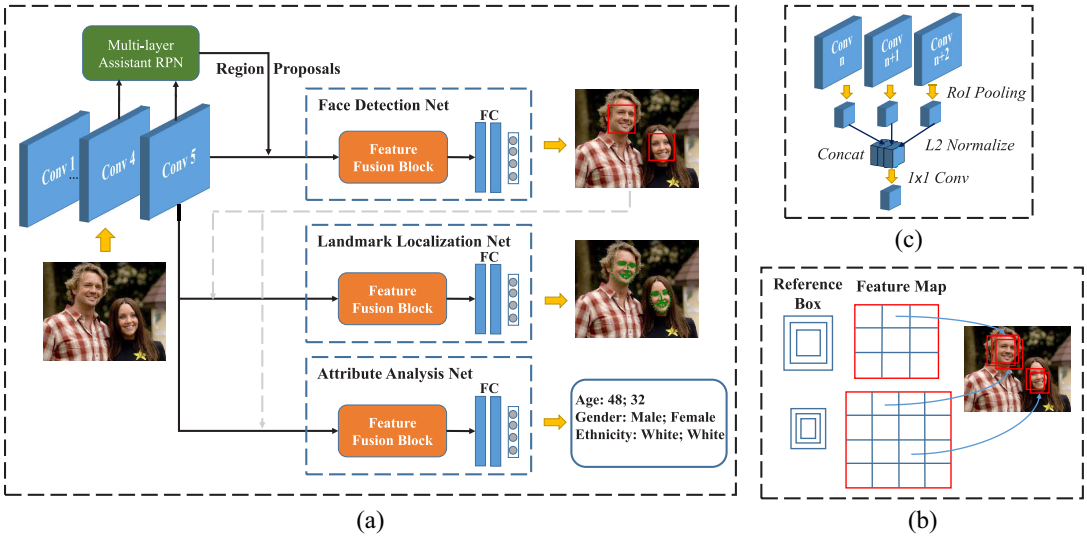


Fig. 3. (a) Given an input image, multilayer assistant RPN first generates numerous coarse region proposals. And then, these proposals are filtered and refined by the face detection net. Finally, according to the position of refined proposals, the landmark location net and attribute analysis net can extract feature maps from task-oriented feature-fused blocks, and predict the facial landmarks and attributes, respectively. (b) With different scales of reference boxes, multilayer assistant RPN can respond to different layers of feature maps to jointly catch various faces. (c) Through the task-oriented feature-fused block, each subnetwork can learn specific feature combination from different convolutional stacks adapted to a specific face task.

feature-fused blocks are set to fuse the feature maps from the top layers of the third, fourth, and fifth convolutional stacks (i.e., conv3_3, conv4_3, and conv5_3 layers), while that in the landmark localization net only considers the feature combination between conv4_3 and conv5_3 layers. The detail of deciding an optimal feature fusion strategy for each task is discussed in Section V-A. At the inference stage, all subnetworks share the basic feature maps from the stem layers, and extract task-specific facial feature maps from associated feature-fused blocks.

B. Multitask Loss Function

In our framework, we mainly consider the optimization of four loss functions: L_{rpn}, L_{det}, L_{land}, and L_{attr} acting on multilayer assistant RPN, face detection net, landmark localization net, and attribute analysis net, respectively. L_{rpn} is used to learn the response for face/nonface decision in each location of given reference boxes, and fit their positions to matched ground-truth boxes. L_{det} is adopted to learn the face/nonface classification for generated region proposals and the corresponding regression of their refined positions. According to the positions of predicted faces, we make a correlative regression for the position of each facial landmark by L_{land}. For attribute analysis, L_{attr} is used to learn three classification tasks, including age (101 classes) estimation, gender (2 classes), and ethnicity (3 classes) recognition. In these functions, L_{cls} and L_{reg} denote softmax loss and smooth L1 loss defined in [43], respectively. With above definitions, we minimize an objective function as follows:

$$L = \sum_i L_{\text{rpn}}(i) + \sum_j L_{\text{det}}(j) + \sum_k L_{\text{land}}(k) + L_{\text{attr}}(k) \tag{1}$$

where i, j, and k are the indexes of a reference box, a proposal object, and a detected face in a mini-batch, respectively. At the training stage, we define a indicator variable I_k^c, where $c \in \{\text{Age, Gender, Ethnicity, Landmark}\}$ to denote the corresponding label of the kth instance. If the c label is missing, I_k^c is 0 and 1 otherwise.

1) RPN Loss: In the original RPN [42], the reference box is assigned a positive label when its IoU overlap is the highest or higher than 0.7 with any ground-truth object. The assignment would hinder the matching of reference boxes especially for the face samples with small sizes. Instead of the above strategy, we choose the reference boxes with top-three highest IoU overlaps as positive samples and minimize the loss function L_{rpn} as follows:

$$L_{\text{rpn}}(i) = \lambda_{\text{cls}} L_{\text{cls}}\left(p_i, p_i^*\right) + \lambda_{\text{reg}} p_i^* L_{\text{reg}}\left(b_i, b_i^*\right) \tag{2}$$

where p_i is the predicted probability of the reference box i being a face. The ground-truth label p_i^* is 1 if the reference box is positive and 0 otherwise. As defined in [42], bi is a vector representing the four parameterized coordinates of the predicted bounding box, and b_i^* is that of the ground truth box associated with a positive reference box. λ_{cls} and λ_{reg} are introduced as the regularization parameters balancing the loss of box classification and box regression.

2) Face Detection Loss: After getting a diverse set of proposals from RPN, we assign them a positive label if the IoU overlap is higher than 0.5 with any ground-truth object and a negative label otherwise. At this stage, the face detection net would be learned to filter the falsely predicted proposals and refine the positions of the truly predicted ones by minimizing the detection loss function L_{det} as follows:

$$L_{\text{det}}(j) = \lambda_{\text{cls}} L_{\text{cls}}\left(d_j, d_j^*\right) + \lambda_{\text{reg}} d_j^* L_{\text{reg}}\left(t_j, t_j^*\right) \tag{3}$$

where d_j and d_j^* denote the predicted probability of the proposal j being a face and the corresponding ground-truth label,

respectively. If the proposal is positive, d_j^* is 1 and 0 otherwise. With the same definition in RPN, t_j is a vector representing the four parameterized coordinates of the predicted face, and t_j^* is that of ground-truth face associated to a positive proposal. $d_j^* L_{\text{reg}}$ means the regression loss is activated only for positive proposal and disabled otherwise.

3) Landmark Localization Loss: Since the position of each facial landmark is closely correlated with that of the detected face, we choose all ground-truth faces and predicted faces with the IoU overlap higher than 0.7 with any ground-truth faces as training samples. The loss function of landmark localization is described as follows:

$$L_{\text{land}}(k) = \lambda_{\text{land}} I_k^{\text{Landmark}} L_{\text{reg}}\left(S_k, S_k^*\right) \tag{4}$$

where S_k and S_k^* are a vector containing 136 parameterized coordinates, which represent the position of 68 predicted landmarks and ground-truth landmarks, respectively. λ_{land} is a regularization parameter of the landmarks localization loss.

For landmarks regression, we refer to the parameterizations of bounding box and adopt parameterized coordinates to denote every landmark point as follows:

$$\begin{aligned} S_k &= (q_1, \ldots, q_{68}) \\ q &= \left(\frac{c_{x_k} - u}{w_k}, \frac{c_{y_k} - v}{h_k}\right) \end{aligned} \tag{5}$$

where q is the two parameterized coordinates of a facial landmark. c_{x_k}, c_{y_k}, w_k, and h_k denote the center coordinates of predicted face k and its width and height, respectively. u and v represent the position of the predicted landmark point.

4) Attribute Analysis Loss: Following the assignment of positive label in landmark localization net, we use the same set of samples to train three facial attribute tasks. The loss function of attribute analysis is defined by

$$\begin{aligned} L_{\text{attr}}(k) = &\lambda_{\text{age}} I_k^{\text{Age}} L_{\text{cls}}\left(a_k, a_k^*\right) \\ &+ \lambda_{\text{gender}} I_k^{\text{Gender}} L_{\text{cls}}\left(g_k, g_k^*\right) \\ &+ \lambda_{\text{ethnicity}} I_k^{\text{Ethnicity}} L_{\text{cls}}\left(r_k, r_k^*\right) \end{aligned} \tag{6}$$

where a_k, g_k, and r_k denote the predicted probability of every class for age with 101 classes, gender with 2 classes, and ethnicity with 3 classes, respectively; while a_k^*, g_k^*, and r_k^* represent the ground-truth labels corresponding to each attribute. The regularization parameter λ_{age}, λ_{gender}, and $\lambda_{\text{ethnicity}}$ are used to balance the loss of each attribute.

C. Alternative Multitask Training Scheme

In multitask learning, it is hard to keep each task work well since some tasks are likely to be overfitting earlier than the others during joint training, which affects the performance of the whole model. To consider different learning capacity of each task in a multitask framework, we design an alternative multitask training scheme containing four stages of optimizations for each task.

1) The proposed network is trained with all tasks in a standard multitask manner. At this stage, we initialize the weights of all stem convolutional layers and fully connected layers by using VGG-16 model. With the same regularization parameters, all task-specific nets share a set of unified training parameters, including learning rate, step size, and number of total iterations, which means each task can be treated fairly but may affect each other due to different learning difficulties.
2) Based on the learned model, we independently optimize RPN and face detection net associated to the performance of detection task. To keep other tasks unaffected, we cut-off the backpropagation of irrelevant layers, including other task-specific nets and their feature-fused blocks. Note that the stem convolutional layers can be still learned for providing different levels of shared feature maps at this stage.
3) After the second stage, the face detection net can generate more accurate face bounding boxes than RPN. We employ these bounding boxes to create training samples with a slight shaking around the detected face. During the training, except for the landmark localization net with associated feature-fused block, the other parts of network are frozen and their weights remain unchanged.
4) With the same training samples and strategy, we optimize the attribute analysis net independently by learning the specific feature combination adapted to the prediction of each attribute, including age estimation, gender, and ethnicity recognition.

Since the alternative training scheme can optimize each task-specific net independently in a multitask network, it not only avoids the potential affect due to different learning difficulty of each task but also benefits from shared feature maps without repeating backward updating. By this way, we can train the proposed TFN on different datasets used for different face tasks and ensure their performances simultaneously. More details and experimental results are provided in Section V-C.

V. Experiment

In this section, we first make a thorough analysis of each task on feature fusion learning. And then, we evaluate the effectiveness of the proposed TFN comparing with several recent multitask frameworks on the JFA dataset. At last, we further assess the performance of our approach on several common benchmarks.

A. Analysis of Feature Fusion Learning

To prove the effectiveness of feature fusion and decide an optimal feature combination for each task, we adopt an increasing top-down feature fusion strategy, in which a task-specific subnetwork will fuse increasing feature maps from the top layers of five convolutional stacks, while other task branches only receive the feature maps from the top convolutional layer. To ensure the fairness of experiment, we initialize the learning rate of all test models to 0.0001 with the reduction of 1/10 every 50k iterations. During the training, we randomly select 50k training samples from the JFA dataset instead of using full training samples, since our purpose is to analyze the effect of different feature fusions for the performance of each task.

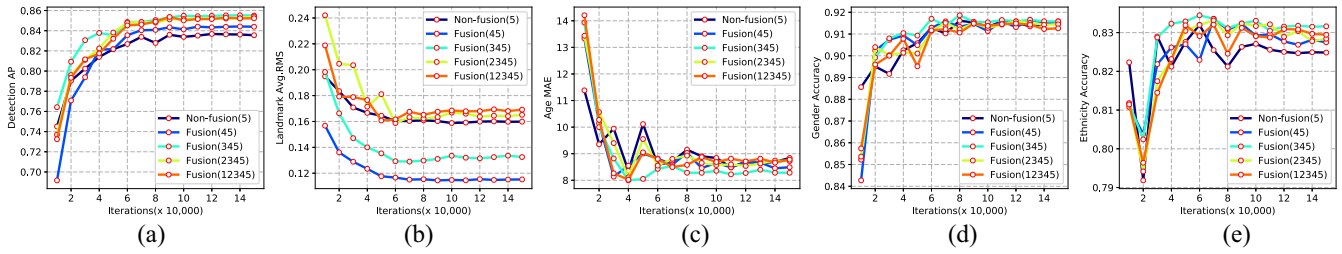


Fig. 4. Effect of different feature fusions for the performance of each task on the JFA testing set. The name of curve denotes a specific feature combination, for example, the Fusion(45) in (a) indicates that the face decision net fuses feature maps from the top layers of the fourth and fifth convolutional stacks. (a) Face detection. (b) Landmark localization. (c) Age estimation. (d) Gender recognition. (e) Ethnicity recognition.

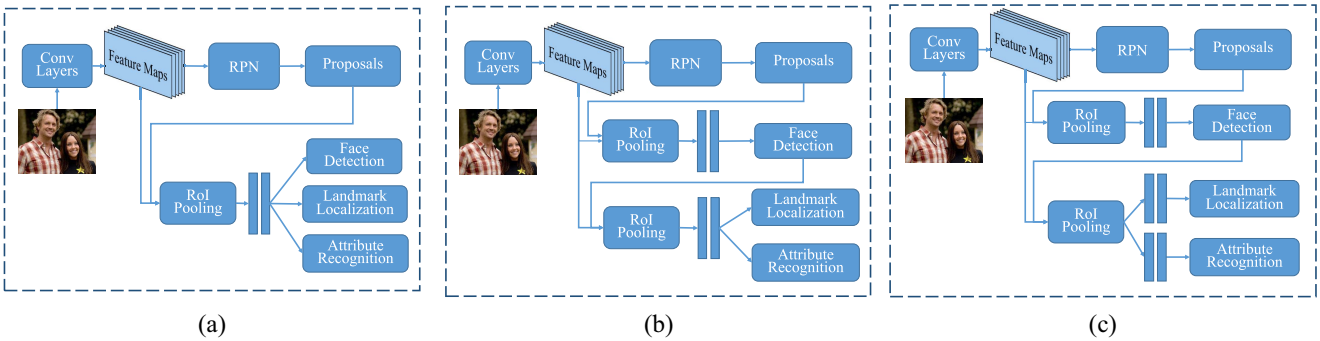


Fig. 5. (a) By adding additional decision layers behind the fully connected layers, the Faster-RCNN can be extended to support multitask prediction. (b) With a new region-wise subnetwork, a two-step refinement is applied into the multitask framework. (c) By introducing a group of unshared fully connected layers, each task can work in an independent decision branch.

As shown in Fig. 4, we can see the variation of performances for different fusion models on the JFA testing set. For the face detection, all the fusion models have higher AP than nonfusion model when they converges. Especially, for the feature combination among the top layers of the third, fourth, and fifth convolutional stacks, it can improve the performance of detection with a large margin. From Fig. 4(b), the Avg.RMS of landmark localization have an obvious decline by exploiting the feature maps from the top-three convolutional stacks. It suggests that these feature maps can facilitate the capacity of location regression in different degrees, where the feature combination between the top layers of the fourth and fifth convolutional stacks can lead to an optimal result. However, the introduction of feature maps from the first and second convolutional stacks dramatically degrades the performance, which may be that the localization task is sensitive for the quite low level of feature maps. For the attribute analysis, we can find that the feature fusion among the top layers of the top-three convolutional stacks more effectively improves the recognition of global attributes than other fusion strategies. All of these experiments indicate that a suitable feature fusion strategy designed in terms of the property of different tasks can produce better performance than using the unified feature maps from the top convolutional layer.

B. Comparison on the JFA Dataset

In this section, we mainly evaluate the proposed model by comparing with several recent joint multitask frameworks, which are based on an extended region convolutional network (i.e., Faster R-CNN [42]). To comprehensively access the performances of all face tasks stated in this paper, we reimplement these methods for joint face analysis as shown in Fig. 5. The ORMM is a standard region-based multitask framework, which have been exploited by the work [33], [34] to joint face/person detection and recognition. By adding an additional decision layer behind the original fully connected layers, a new task can be easily introduced into the framework. However, since the accuracy of detected boxes from RPN is limited, these follow-up tasks may receive unreliable feature maps. The TRMM is a coarse-to-fine multitask framework with two-step refinement, which have been applied into joint vehicle analysis [32]. The main difference from ORMM is exploiting a two-step refinement of detected boxes by adding a new region-wise subnetwork. The SRMM is a variant of TRMM with task-specific fully connected layers. By introducing a group of unshared fully connected layers, a new task can work in an independent decision branch. In the proposed TFN, the SRMM is extended by replacing the original RPN with the multilayer assistant RPN, and adding task-oriented feature-fused block in each decision branch as shown in Fig. 3(a). To ensure a fair comparison, we standardized the base network architectures (i.e., VGG-16) amongst all the methods, and train them on full training set from JFA with the same training setting, in which the learning rate of each model is initialized to 0.0001, and gradually decreased by 1/10 every 100k iterations until the model converges at 300K iterations.

From Table III, five task-specific evaluation metrics are reported for different multitask frameworks on JFA testing set. In addition, Fig. 6 describes the performances of face detection and landmark localization by showing the precision–recall curves and cumulative error curves with corresponding

TABLE III
COMPARISONS BETWEEN THE PROPOSED FRAMEWORK AND OTHER MULTITASK METHODS ON THE JFA TESTING SET

Method	Detection AP↑	Landmark Avg.RMS↓	Age MAE↓	Gender Accuracy↑	Ethnicity Accuracy↑
ORMM	0.8640	0.6602	8.454	91.18%	82.34%
TRMM	0.8572	0.1755	8.560	91.25%	82.12%
SRMM	0.8628	0.1605	8.465	91.26%	82.37%
Ours	**0.9323**	**0.0915**	**7.580**	**92.09%**	**83.67%**

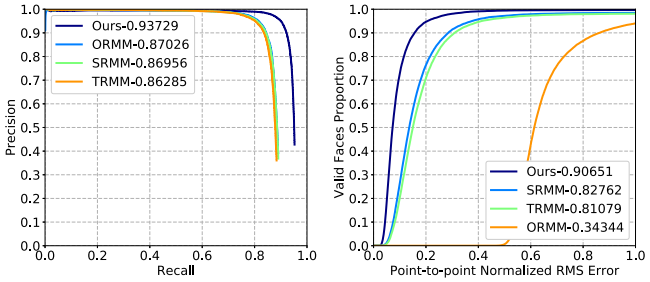


Fig. 6. Accuracy of detection and landmark localization for all the methods is reported as (a) precision–recall curves and (b) cumulative error curves on the JFA testing set.

AUC values, in which the larger the AUC is, the better the performance of model is. We can see that the methods with two-step refinement of bounding box can effectively facilities the performance of landmark localization. It means that the landmarks regression is much benefit from the accurate location of face box, since the localization task relies on accurate feature maps, and the position of landmarks is closely correlated with valid face region in terms of the definition of parameterized landmarks in Section IV-B. Moreover, by introducing independent decision branches, the performance of landmark localization and attributes recognition in SRMM is slightly better than TRMM, perhaps because the unshared fully connected layers can reduce the impact among these tasks when joint training. With the extended RPN and task-oriented feature-fused block, the proposed TFN can effectively improve the performance of face detection task. Benefited from the improved accuracy of detected face and task-specific feature combinations, the performances of landmark localization and attributes recognition can also be enhanced further.

C. Comparison on Common Face Benchmarks

To validate the generalization of the proposed TFN, we compare our model with recent state-of-the-art methods on three common face benchmarks, that is, WIDER FACE [25], 300W [26], and MORPH II [24].

1) WIDER FACE: The dataset have been widely used to validate the performance of face detector like [35], [45], [46], since it provides three levels of validation and testing sets covering various faces with different scales, poses, and occlusions. It contains 32 203 images with 393 703 annotated faces, in which they are randomly divided into three subsets for training, validation, and testing by 40%/10%/50% proportions.

2) 300W: As shown in Table I, since most of the existing face datasets like LFPW [19], HELEN [20], and AFW [21] are annotated with different numbers of landmarks, 300W [26] presents a unified protocol to reannotate these datasets as the training set with 68 landmarks. In addition, it provides 300 indoor and 300 outdoor face images as testing set to evaluate the performance of landmark localization.

3) MORPH II: The dataset contains about 55 000 face images with precise age, gender, and ethnicity labels, and can be used to train and test multitask models for age estimation, gender, and ethnicity recognition. Following the works [39], [47], it can be split into three nonoverlapped subsets (i.e., S1, S2, and S3) for crossed training and testing, in which there are about 10 634 images in the training set and about 44 610 images are used for testing.

In this experiment, we train two models with or without the pretraining of the JFA dataset (i.e., TFN+JFA or TFN) in different common dataset to evaluate the contribution of the proposed TFN and JFA dataset. According to the alternative training scheme stated in Section IV-C, we first use the training set provided by WIDER FACE [25] to optimize face detection net. In this process, the initial learning rate is set to 0.001, and dropped to 1/10 every 50k iterations until the number of total iterations is 150k. And then, with the initial learning rate 0.01 and the same learning strategy, landmark localization net is trained based on reannotated LFPW [19], HELEN [20], and AFW [21] datasets. Finally, we optimize the attribute analysis net by exploiting MORPH II [24] training set with the same learning parameters as those of landmark localization net. During training, all regularization parameters are set to 1 for training each task fairly. Moreover, all training images are horizontally flipped and expanded with three scale ratios {0.5, 1, 2}. And no other data augmentation is used.

From Fig. 7, we can see that the TFN achieves competitive performance on WIDER FACE [25] validation and testing sets comparing with the state-of-the-art face detectors, such as [13], [35], [45], [46], and [55]–[61]. Especially, with the pretraining on the JFA dataset, the AUCs can be increased with an obvious margin, which outperform most of the recent methods on the easy and medium sets. It implies that the proposed network can effectively learn discriminative representations from large training samples covering various faces provided by the JFA dataset. However, compared to the result of state-of-the-art method on the hard set, our method is relatively weak even if with the pretraining of the JFA dataset. The result is expected since the hard set contains many difficultly recognizable faces with tiny scale and heavy occlusion, which are not mainly concerned by this paper. In the JFA dataset, these faces with missing information and cannot be annotated for facial landmarks and attributes. In fact, in contrast to recent face detectors, our approach can better detect the recognizable faces, which are more valuable in real-world applications than tiny or heavily occlusive faces.

Fig. 8 shows the cumulative error curves of the proposed TFN and all the participant methods [48]–[52] reported in the second 300W competition [26]. We can see that the TFN get competitive performance with 78.43% and 78.42% AUCs on the indoor and outdoor datasets, respectively. With the pretraining on the JFA dataset, the TFN can further promote the accuracy of predicted landmarks in different scenarios.

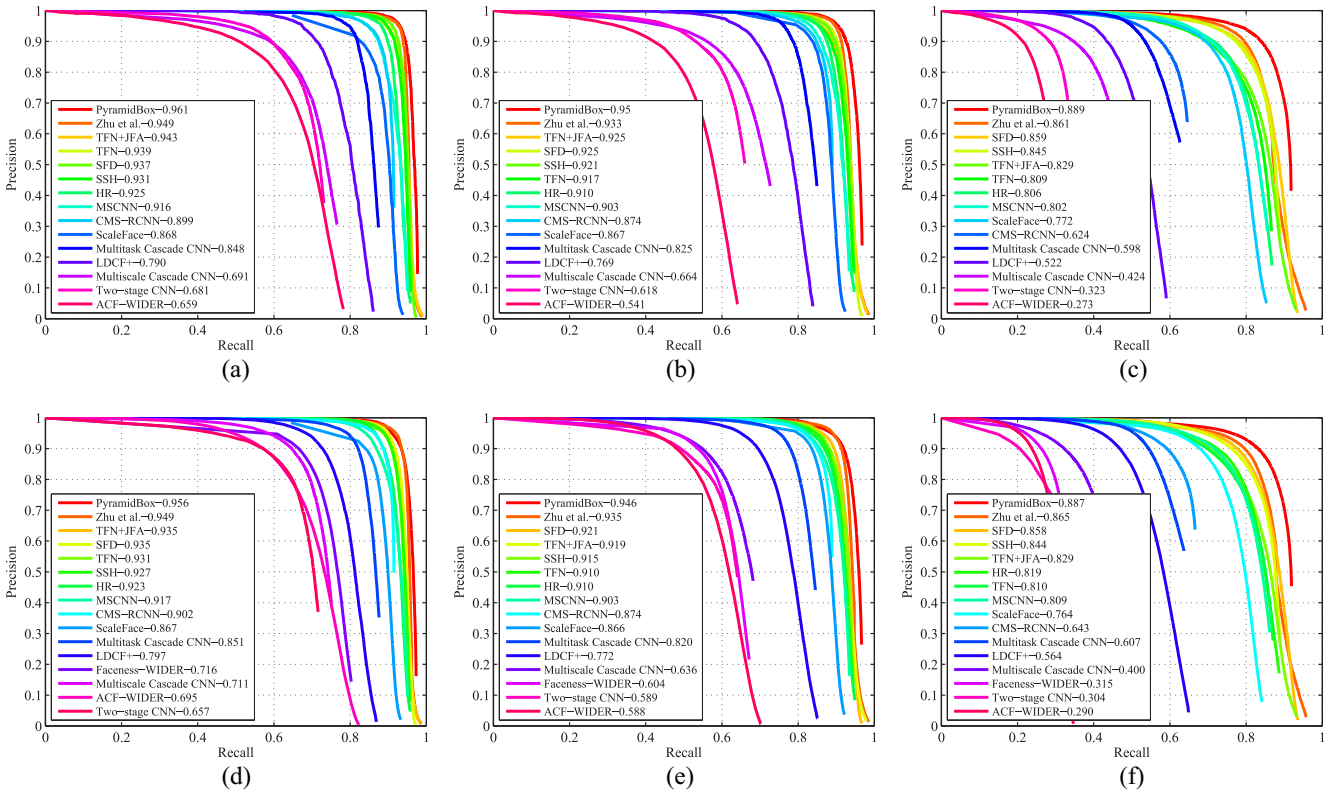


Fig. 7. Comparisons between the proposed framework and other methods on WIDER FACE validation and testing sets. (a) Val: Easy. (b) Val: Medium. (c) Val: Hard. (d) Test: Easy. (e) Test: Medium. (f) Test: Hard.

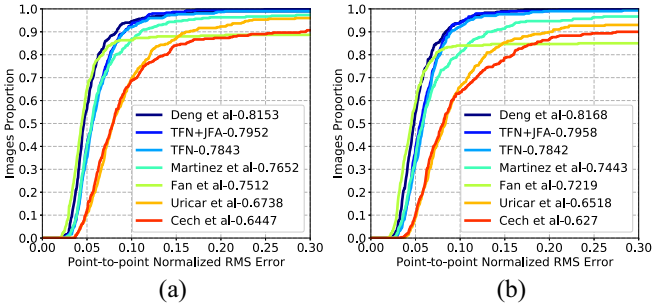


Fig. 8. Comparison of the proposed framework with other methods on the 300W dataset (indoor + outdoor). (a) Indoor, 68 points. (b) Outdoor, 68 points.

TABLE IV
COMPARISONS BETWEEN THE PROPOSED FRAMEWORK AND OTHER METHODS ON THE 300W DATASET

Method	Max Error=0.1		Max Error=0.2		Max Error=0.3	
	AUC↑	FR↓	AUC↑	FR↓	AUC↑	FR↓
Uricar et al. [48]	0.2109	32.17%	0.5257	8.83%	0.6628	5.50%
Cech et al. [49]	0.2218	33.83%	0.5087	13.00%	0.6359	9.67%
Martinez et al. [50]	0.3779	16.00%	0.6504	4.50%	0.7547	3.00%
Deng et al. [51]	0.4752	5.50%	**0.7274**	0.67%	**0.8160**	0.67%
Fan et al. [52]	0.4802	14.83%	0.6710	13.50%	0.7366	13.17%
DenseReg [53]	0.3605	10.83%	-	-	-	-
DenseReg+MDM [53]	0.5219	3.67%	-	-	-	-
LAB [54]	**0.5885**	**0.83%**	-	-	-	-
TFN	0.3932	7.83%	0.6813	1.17%	0.7843	0.83%
TFN+JFA	0.4076	6.83%	0.6946	**0.50%**	0.7955	**0.17%**

Table IV shows the newest results, including the AUC and failure rate (FR) of the recent state-of-the-art methods [53], [54] on the 300W test set. We can find that the performance of TFN is relatively weak comparing with the state-of-the-art results especially for the maximum error of 0.1, while the AUCs and FRs of our method are improved significantly for the maximum errors of 0.2 and 0.3. It means that the TFN can keep a high recall rate for detected faces with gradually increasing point-to-point RMS error, but cannot ensure the highly accurate fitting for the locations of all the landmark points. In Section VI, we will discuss the optimization of single task in the proposed TFN and detail an extension of the landmark localization net for improving the accuracy of predicted landmarks without affecting the performances of other tasks.

Table V reports the performances of several recent methods [39]–[41], [62]–[64] for the predictions of age, gender, and ethnicity on MORPH II [24]. We can see that the TFN achieves competitive performance for all the tasks comparing to the state-of-the-art work. By adopting the pretraining on the JFA dataset, the Avg.E of age can be obviously reduced by about 0.2, while the accuracies of gender and ethnicity are slightly increased and outperform the latest results on [64]. It is worth mentioning that the GenderRace2Age and RaceGender2Age methods [64] cascade three VGG-16 nets for three task-specific inferences, while the TFN model only exploits the partial parameters of a VGG-16 net by introducing a shallow attribute analysis net for all the tasks.

TABLE V
COMPARISONS BETWEEN THE PROPOSED FRAMEWORK AND OTHER METHODS ON THE MORPH II DATASET

Method	Training Set	Testing Set	Age		Gender	Ethnicity
			MAE↓	Avg.E↓	Accuracy↑	Accuracy↑
BIF+KPLS	S1	S2+S3	4.21	4.18	98.20%	98.90%
[41]	S2	S1+S3	4.15		98.20%	98.80%
BIF+KCCA	S1	S2+S3	4.01	3.98	98.50%	99.00%
[40]	S2	S1+S3	3.95		98.40%	99.00%
Multi-scale	S1	S2+S3	3.72	3.63	98.00%	98.65%
CNN [39]	S2	S1+S3	3.54		97.80%	98.55%
HiGSFA [62]	S1	S2+S3	3.51	3.497	97.71%	**99.16%**
	S2	S1+S3	3.49		97.69%	**99.14%**
Compact	S1	S2+S3	3.22	3.23	98.84%	-
CNN [63]	S2	S1+S3	3.25		98.79%	-
GenderRace2Age	S1	S2+S3	**3.143**	**2.99**	98.23%	97.78%
[64]	S2	S1+S3	2.839		98.70%	97.99%
RaceGender2Age	S1	S2+S3	3.145	**2.99**	98.23%	97.78%
[64]	S2	S1+S3	**2.838**		98.70%	97.99%
TFN	S1	S2+S3	3.23	3.205	98.71%	98.05%
	S2	S1+S3	3.18		98.66%	98.13%
TFN+JFA	S1	S2+S3	3.15	3.045	**99.07%**	98.35%
	S2	S1+S3	2.94		**99.06%**	98.47%

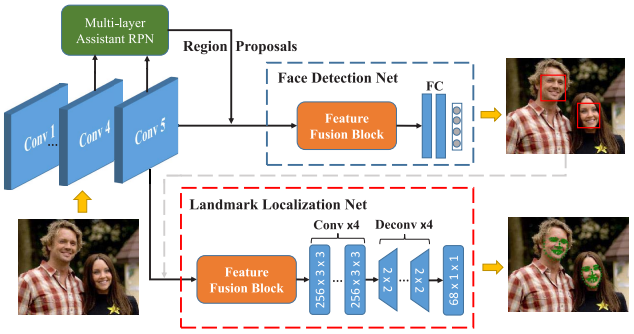


Fig. 9. Red dotted box shows a new landmark localization net with the encode–decode structure for the landmark heatmap classification.

TABLE VI
COMPARISONS BETWEEN THE PROPOSED TFN AND ITS IMPROVED VERSIONS ON THE 300W DATASET

Method	Max Error=0.1		Max Error=0.2		Max Error=0.3	
	AUC↑	FR↓	AUC↑	FR↓	AUC↑	FR↓
TFN	0.3932	7.83%	0.6813	1.17%	0.7843	0.83%
TFN+JFA	0.4076	6.83%	0.6946	0.50%	0.7955	**0.17%**
TFN+JFA (Heatmap)	0.4483	3.83%	0.7188	0.33%	0.8114	0.33%
TAFN+JFA (Heatmap)	**0.4796**	**3.17%**	**0.7356**	**0.33%**	**0.8226**	0.33%

VI. DISCUSSION

A. Optimization of Single Task in TFN

In this paper, we mainly propose a multitask CNN scheme for jointly solving different face tasks. For the follow-up work, a natural thought is to optimize the single task inside the framework without affecting the performance of other tasks. Since the TFN has a completely modular design, each task-specific subnetwork with corresponding feature-fused block is decoupled from other parts of this network and can be optimized independently. As an example, we further improve the performance of facial landmark localization by designing a new landmark localization net in the following.

Inspired by the recent studies [53], [54] for the landmark localization by solving a heatmap prediction problem, we modify the landmark localization net in support of the landmark heatmap classification. As shown in Fig. 9, the input feature maps from the feature-fused block are first encoded four times by the consecutive convolution operations with 3×3 kernel size. And then, through the learned deconvolution operations with 2×2 kernel size, the feature maps are gradually decoded with 16 times larger size (112×112) than the input ones. Finally, we can get 68 landmark heatmaps representing the response location of each facial landmark by using a 1×1 convolution filter. In this scheme, we adopt a per-pixel softmax and a multinomial cross-entropy loss function defined in [65]. Benefited from the alternative training strategy, we do not have to retrain the entire network and only fine-tune the landmark localization net with a new initial learning rate (0.001). From Table VI, TFN+JFA (Heatmap) denotes the modified TFN with landmark heatmap prediction. We can find that the AUCs of cumulative error curves are significantly increased especially for the maximum error of 0.1, which indicates heatmap prediction is more suitable for the landmark localization task than coordinate regression.

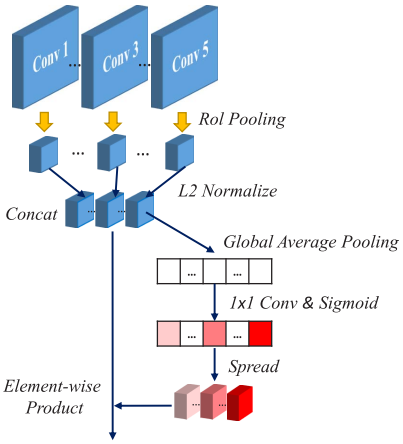


Fig. 10. Extension of task-oriented feature-fused block with the squeeze-and-excitation operation.

B. Task-Oriented Automatic Feature Fusion

In the TFN, we design a task-oriented feature-fused block for learning the feature combination from the top layers of selected convolutional stacks which are determined by an increasing top-down feature fusion strategy. There are two potential drawbacks in this scheme.

1) The selection of feature fusion layers is time consuming since it will be repeatedly executed when considering a new task.
2) This scheme is based on a strong assumption that only selected layers are available for a specific task, which may lose some information that is helpful but not required. In the following, we develop an initial scheme for improving the original feature-fused block in support of the automatic feature fusion.

Based on the work [66], we add a simplified squeeze-and-excitation operation into the task-oriented feature-fused block as shown in Fig. 10. First, the feature maps from the top layers of all the convolutional stacks are input to the block

to generate a set of normalized region-wise features. And then, through the global average pooling, each feature map is squeezed into a channel-wise statistic value to describe the global spatial information. Finally, these statistic values are further encoded by a 1×1 convolution filter and activated by a sigmoid operation. The activated vector represents the significance degrees of input feature maps and is used for their element-wise reweighting. As an example, we apply the new block into the modified landmark localization net called TAFN+JFA (Heatmap). From Table VI, we can see that the AUCs for different maximum errors are further improved by about 1%–3%. It suggests the effectiveness of the new block for mining the potential information from different convolution layers. A future research direction is to study the generalization of automatic feature fusion and the sparsity constraint on the weight vector.

VII. Conclusion

Compared to the pipeline model with redundant intermediate processing, the multitask method based on CNN is more promising since it can make an inference in an end-to-end fashion, in which each task can be jointly optimized by minimizing the multitask loss function. In this paper, we indicate several main restraints of developing the multitask model for joint face analysis, and propose a new multitask network as well as a multivariate face dataset. In this network, we design task-oriented feature-fused blocks, in which each task can learn suitable feature combination for robust prediction. Moreover, we present an alternate learning scheme used to reduce the potential impacts among tasks when joint learning. From our experiments, we make a thorough analysis for all tasks on feature fusion learning and decide their optimal fusion strategy. By comparing with several recent multitask models, we show the effectiveness of the TFN and four benchmark methods on the new dataset for further investigation. Finally, several extended experiments also suggest that the TFN with JFA dataset have robust generalization ability and can simultaneously get promising results on three common face benchmarks.

References

[1] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2015, pp. 435–442.

[2] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017.

[3] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 34–42.

[4] A. Dehghan, E. G. Ortiz, G. Shu, and S. Z. Masood, "DAGER: Deep age, gender and emotion recognition using convolutional neural network," *arXiv preprint*, 2017. [Online]. Available: https://arxiv.org/abs/1702.04280

[5] Y. Huang, W. Wang, L. Wang, and T. Tan, "Conditional high-order Boltzmann machines for supervised relation learning," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4297–4310, Sep. 2017.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[8] Y. Huang, W. Wang, and L. Wang, "Unconstrained multimodal multi-label learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1923–1935, Nov. 2015.

[9] C. Zhang and Z. Zhang, "Improving multiview face detection with multi-task deep convolutional neural networks," in *Proc. IEEE Win. Conf. Appl. Comput. Vis.*, 2014, pp. 1036–1041.

[10] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 94–108.

[11] M. Ehrlich, T. J. Shields, T. Almaev, and M. R. Amer, "Facial attributes classification using multi-task representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 47–55.

[12] 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," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

[13] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[14] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "DenseBox: Unifying landmark localization with end to end object detection," *arXiv preprint*, 2015. [Online]. Available: https://arxiv.org/abs/1509.04874

[15] Y. Li, B. Sun, T. Wu, and Y. Wang, "Face detection with end-to-end integration of a ConvNet and a 3D model," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 420–436.

[16] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Image Vis. Comput.*, vol. 32, no. 10, pp. 790–799, 2014.

[17] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," Univ. Massachusetts Amherst, Amherst, MA, USA, Rep. UM-CS-2010-009, 2010.

[18] B. F. Klare *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1931–1939.

[19] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2930–2940, Dec. 2013.

[20] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 679–692.

[21] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2879–2886.

[22] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 768–783.

[23] R. Rothe, R. Timofte, and L. Van Gool, "DEX: Deep expectation of apparent age from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 10–15.

[24] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. 7th IEEE Int. Conf. Automat. Face Gesture Recognit.*, 2006, pp. 341–345.

[25] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5525–5533.

[26] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2013, pp. 397–403.

[27] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Fine-grained evaluation on face detection in the wild," in *Proc. 11th IEEE Int. Conf. Workshops Automat. Face Gesture Recognit.*, vol. 1, 2015, pp. 1–7.

[28] M. Köestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 2144–2151.

[29] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio Video Based Biometric Person Authentication*, vol. 964, 1999, pp. 965–966.

[30] J. R. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[32] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teulière, and T. Chateau, "Deep MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2040–2049.

[33] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3415–3424.

[34] K. He, Y. Fu, and X. Xue, "A jointly learned deep architecture for facial attribute analysis and face detection in the wild," *arXiv preprint*, 2017. [Online]. Available: https://arxiv.org/abs/1707.08705

[35] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3FD: Single shot scale-invariant face detector," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 192–201.

[36] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 787–796.

[37] S. Escalera *et al.*, "Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 1–9.

[38] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[39] D. Yi, Z. Lei, and S. Z. Li, "Age estimation by multi-scale convolutional network," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 144–158.

[40] G. Guo and G. Mu, "Joint estimation of age, gender and ethnicity: CCA vs. PLS," in *Proc. 10th IEEE Int. Conf. Workshops Automat. Face Gesture Recognit.*, 2013, pp. 1–6.

[41] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 657–664.

[42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[43] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[44] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," in *Proc. Int. Conf. Learn. Represent. Workshops*, 2016. [Online]. Available: https://arxiv.org/abs/1506.04579

[45] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "SSH: Single stage headless face detector," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4875–4884.

[46] P. Hu and D. Ramanan, "Finding tiny faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 951–959.

[47] Z. Tan, S. Zhou, J. Wan, Z. Lei, and S. Z. Li, "Age estimation based on a single network with soft softmax of aging modeling," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 203–216.

[48] M. Uřičář, V. Franc, D. Thomas, A. Sugimoto, and V. Hlaváč, "Multi-view facial landmark detector learned by the structured output SVM," *Image Vis. Comput.*, vol. 47, pp. 45–59, Mar. 2016.

[49] J. Čech, V. Franc, M. Uřičář, and J. Matas, "Multi-view facial landmark detection by using a 3D shape model," *Image Vis. Comput.*, vol. 47, pp. 60–70, Mar. 2016.

[50] B. Martinez and M. F. Valstar, "$L_{2,1}$-based regression and prediction accumulation across views for robust facial landmark detection," *Image Vis. Comput.*, vol. 47, pp. 36–44, Mar. 2016.

[51] J. Deng, Q. Liu, J. Yang, and D. Tao, "M^3 CSR: Multi-view, multi-scale and multi-component cascade shape regression," *Image Vis. Comput.*, vol. 47, pp. 19–26, Mar. 2016.

[52] H. Fan and E. Zhou, "Approaching human level facial landmark localization by deep learning," *Image Vis. Comput.*, vol. 47, pp. 27–35, Mar. 2016.

[53] R. A. Guler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos, "DenseReg: Fully convolutional dense shape regression in-the-wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6799–6808.

[54] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2129–2138.

[55] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," in *Proc. IEEE Int. Joint Conf. Biometrics*, 2014, pp. 1–8.

[56] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3676–3684.

[57] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection," in *Deep Learning for Biometrics*. Cham, Switzerland: Springer, 2017, pp. 57–79.

[58] E. Ohn-Bar and M. M. Trivedi, "To boost or not to boost? On the limits of boosted trees for object detection," in *Proc. 23rd IEEE Int. Conf. Pattern Recognit.*, 2016, pp. 3350–3355.

[59] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 354–370.

[60] C. Zhu, R. Tao, K. Luu, and M. Savvides, "Seeing small faces from robust anchor's perspective," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5127–5136.

[61] X. Tang, D. K. Du, Z. He, and J. Liu, "PyramidBox: A context-assisted single shot face detector," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 797–813.

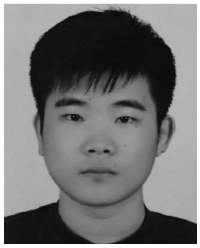
[62] A. Escalante and L. Wiskott, "Improved graph-based SFA: Information preservation complements the slowness principle," *arXiv preprint*, 2016. [Online]. Available: https://arxiv.org/abs/1601.03945

[63] Y. Yang *et al.*, "Video system for human attribute analysis using compact convolutional neural network," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 584–588.

[64] J. Wan, Z. Tan, Z. Lei, G. Guo, and S. Z. Li, "Auxiliary demographic information assisted age estimation with cascaded structure," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2531–2541, Sep. 2018.

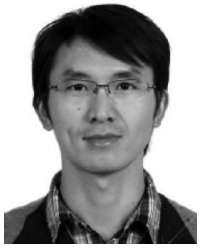
[65] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[66] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.



Xuxin Lin received the B.S. degree in software engineering from the Beijing Institute of Technology, Zhuhai, China, in 2014 and the M.S. degree from the Macau University of Science and Technology, Macau, China, in 2016, where he is currently pursuing the Ph.D. degree.

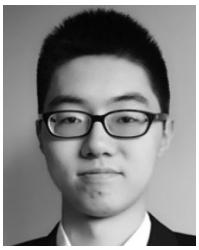
His current research interests include computer vision and pattern recognition.



Jun Wan (M'16) received the B.S. degree from the China University of Geosciences, Beijing, China, in 2008 and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, in 2015.

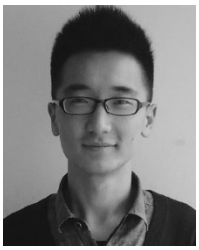
Since 2015, he has been an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. He has published papers in top journals, such as the *Journal of Machine Learning Research*, the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CYBERNETICS, and the *ACM Transactions on Multimedia Computing, Communications, and Applications*. His current research interests include computer vision, and machine learning, especially for gesture and action recognition and facial attribution analysis (i.e., age estimation, facial expression, gender, and race classification).

Dr. Wan has served as the Reviewer on several top journals and conferences, such as the *Journal of Machine Learning Research*, the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, *Pattern Recognition*, ICPR2016, CVPR2017, ICCV2017, and FG2017.



Yiliang Xie is currently pursuing the undergraduation degree in software engineering with the Macau University of Science and Technology, Macau, China.

His current research interests include object detection and face detection.



Shifeng Zhang received the B.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2015. He is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

His current research interests include computer vision and pattern recognition, especially for object detection, face detection, and pedestrian detection.

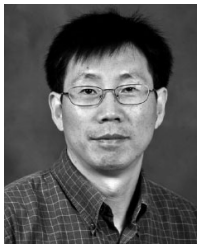
Chi Lin received the B.S. degree (First Class Hons.) from the Faculty of Information Technology, Macau University of Science and Technology, Macau, China, in 2017. He is currently pursuing the M.S. degree with the University of Southern California, Los Angeles, CA, USA.

From 2015 to 2016, he was selected to participate in "Stars of Tomorrow Internship Program" in Microsoft Research Asia, Beijing, China. Since 2017, he has been a Research Intern with the Institute of Automation, Chinese Academy of Sciences, Beijing. His current research interests include machine learning, computer vision, and gesture recognition.

Yanyan Liang (M'08) received the B.S. degree from the Chongqing University of Communication and Posts, Chongqing, China, in 2004 and the M.S. and Ph.D. degrees from the Macau University of Science and Technology (MUST), Macau, China, in 2006 and 2009, respectively.

He is currently an Assistant Professor with MUST. He has published over 30 papers related to pattern recognition, image processing, and computer version. He is also researching on smart city applications with computer vision. His current research interests include computer vision, image processing, and machine learning.

Dr. Liang is currently a Reviewer of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and IEEE ACCESS.



Guodong Guo (M'07–SM'07) received the B.E. degree in automation from Tsinghua University, Beijing, China, and the Ph.D. degree in computer science from the University of Wisconsin–Madison, Madison, WI, USA.

He is currently the Deputy Head of the Institute of Deep Learning, Baidu Research, Beijing, China, and also an Associate Professor with the Department of Computer Science and Electrical Engineering, West Virginia University (WVU), Morgantown, WV, USA. He visited and worked in several places, including INRIA, Sophia Antipolis, France; Ritsumeikan University, Kyoto, Japan; and Microsoft Research, Beijing, China. He has authored a book entitled *Face, Expression, and Iris Recognition Using Learning-Based Approaches* in 2008, co-edited two books entitled *Support Vector Machines Applications* in 2014 and *Mobile Biometrics* in 2017, and published over 100 technical papers. His current research interests include computer vision, biometrics, machine learning, and multimedia.

Dr. Guo was a recipient of the North Carolina State Award for Excellence in Innovation in 2008, the Outstanding Researcher (2017–2018 and 2013–2014) at CEMR, WVU, and the New Researcher of the Year (2010–2011) at CEMR, WVU. He was selected the "People's Hero of the Week" by BSJB under Minority Media and Telecommunications Council in 2013. Two of his papers were selected as "The Best of FG'13" and "The Best of FG'15," respectively.

Stan Z. Li (F'09) received the B.Eng. degree from Hunan University, Changsha, China, the M.Eng. degree from the National University of Defense Technology, Changsha, China, and the Ph.D. degree from Surrey University, Guildford, U.K.

He is currently a Professor and the Director of the Center for Biometrics and Security Research, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He was an Associate Professor with Nanyang Technological University, Singapore. He was with Microsoft Research Asia, Beijing, as a Researcher from 2000 to 2004. He has published over 200 papers in international journals and conferences, and authored and edited eight books. His current research interests include pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance.

Dr. Li was an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and is acting as the Editor-in-Chief of the *Encyclopedia of Biometrics*. He served as the Program Co-Chair for the International Conference on Biometrics in 2007 and 2009, and has been involved in organizing other international conferences and workshops in the fields of his research interest. He was elevated to an IEEE fellow for his contributions to the fields of face recognition, pattern recognition, and computer vision. He is a member of the IEEE Computer Society.