IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 7, 2019, accepted March 18, 2019, date of publication April 4, 2019, date of current version April 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2909327

Head Pose Estimation in the Wild Assisted by Facial Landmarks Based on Convolutional Neural Networks

JIAHAO XIA, LIBO CAO, GUANJUN ZHANG, AND JIACAI LIAO
State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, Hunan University, Changsha 410006, China
Corresponding author: Guanjun Zhang (zgjhuda@163.com)

ABSTRACT Convolutional neural networks (CNNs) exhibit excellent performance on the head pose estimation problem under controllable conditions, but their generalization ability in the wild needs to be improved. To address this issue, we propose an approach involving the introduction of facial landmark information into the task simplifier and landmark heatmap generator constructed before the feed-forward neural network, which can use this information to normalize the face shape into a canonical shape and generate a landmark heatmap based on the transformed facial landmarks to assist in feature extraction, for enhancing generalization ability in the wild. Our method was trained on 300W-LP and tested on AFLW2000-3D. The result shows that for the same feed-forward neural network when our method is used to introduce facial landmark information into a CNN, accuracy improves from 88.5% to 99.0% and mean average error decreases from 5.94° to 1.46° on AFLW2000-3D. Furthermore, we evaluate our method on several datasets used for pose estimation and compare the result with AFLW2000-3D, finding that the features extracted by a CNN could not reflect the head pose efficiently, which limits the performance of the CNN on the head pose estimation problem in wild. By introducing facial landmarks, the CNN could extract features that reflect head pose more efficiently, thereby significantly improving the accuracy of head pose estimation in the wild.

INDEX TERMS Head pose estimation, convolutional neural network, landmark heatmap, faceshape normalization.

I. INTRODUCTION

Image-based head pose estimation is a challenge in the filed of machine vision. During the last decade, various applications, which are based on accurate head pose estimation, have been developed, such as intelligent meeting system [1], human robot interaction [2], person tracking [3], and driver monitor system [4]–[8]. The performance of such applications depends to a large extent on the accuracy and robustness of the head pose estimation algorithm. At present, though the head pose estimation algorithm exhibit excellent performance under controllable conditions, it does not show ideal performance in a wild environment (i.e., when there is large variation in appearance and environmental conditions) [9]. Therefore, improving accuracy and robustness of the head pose estimation algorithm is of considerably significant.

'Head pose' is the head's relative orientation with respect to the camera, and it is typically expressed in term of three angles (yaw, pitch and roll). To predict these angles from a raw image, many methods have been proposed from different points of view, from model-based methods (geometric models [10]–[13] and deformable models [14], [15]) to appearance-based methods (classification approaches [16]–[18], manifold embedding approaches [19]–[22] and nonlinear regression approaches [9], [23], [24]). Model-based methods use geometric information or landmark locations to estimate the head pose, while model-based methods exhibit excellent performance in small angles; the result of estimation relies entirely on landmarks detection performance and the information of the image is ignored, resulting in fragile robustness [25]. Appearance-based methods estimate head pose directly from the raw image. As they take advantage of image information, appearance-based methods are less sensitive to partial occlusions and extreme angular views. Nevertheless, the image contains several features and the output variables that regressed from the features do not necessarily correspond to the pose angle [22]; therefore appearance-based methods have considerable potential to make further progress in estimation accuracy.

The associate editor coordinating the review of this manuscript and approving it for publication was Fatih Emre Boran.

In view of the limitations of these two methods, a new method based on CNN [26], [27] is proposed to improve the accuracy and robustness of head pose estimates in the wild by combining the advantages of these two kinds of methods to complement each other. We use state-of-the-art landmark detector (Dlib [28] and FAN [29]) to find facial landmarks from the image, and then input the landmarks location and image to the task simplifier for calculations. The task simplifier is responsible for calculating the affine transformation matrix that minimizes the standard deviation between input landmarks S_1 and canonical shape S_0. Then the input image is transformed based on the matrix to simplify the head pose estimation task and improve accuracy. Landmark heatmap is generated in the heatmap generator based on the transformed landmarks. Then, the landmark heatmap is stacked with the transformed image to be input into the feed-forward neural network. Thanks to the introduction of landmark heatmap, feed-forward neural network can focus on the area around facial landmarks while extracting features from the image. Compared to appearance-based methods, the proposed method has reduced interference from wild environment (large variation in appearance and environmental conditions) to features extraction, so the output of convolution layers can express the head pose more efficiently and the accuracy of the algorithm is improved. Because the landmarks location does not determine the estimation result directly, the proposed method is more robust than model-based methods.

The main research content and contributions of this study are as follows:

- A new method based on a convolutional neural network for head pose estimation in the wild is proposed, which combines the advantages of model-based methods and appearance-based methods. The impacts of the new architectures proposed in this study have been investigated and it has been proved that these new architectures contribute a lot to the accuracy and robustness of head pose estimation in the wild.
- The generalization ability of the proposed method is demonstrated by validation the method on several datasets in the wild and compared with state-of-the-art methods. The head pose estimation method proposed in this paper reduces the *mean average error* (MAE) from 5.94° to 1.46° and improves accuracy from 88.55% to 99.00% on validation set.
- By comparing the evaluation result under controllable conditions with the result under wild conditions, the main factor which causes a large difference in performance between controllable conditions and wild conditions is found. This is because under controllable conditions, the background changes less and the features extracted by CNN can express head pose more efficiently. The main factor which limites the accuracy of controllable condition datasets is the expressivity of CNN.
- We add facial landmarks annotation for BIWI and CASPEAL datasets by FAN and manually relabel the image with large error, constructing a facial landmark and pose dataset containing 35K images. The dataset will be published later for research at https://github.com/shallybrown/facial_landmark_label-

The rest of this paper is organized as follows. In Section 2, we provide a brief overview of the literature related to head pose estimation. In Section 3, we elaborate on the details of the proposed method. In Section 4, we evaluate the performance of our method on wild datasets as well as controllable condition datasets. Lastly, section 5 presents the study's conclusions.

II. RELATED WORKS

Currently, the head pose estimation problem has been investigated from various perspectives and with various techniques, such as laser pointers, camera arrays, stereo-cameras and magnetic and inertia sensors [30]. Compared to other methods, the methods based on raw images are more feasible and less limited. This section is limited to the methods based on raw images, which are the most relevant methods to our work. A completed description of all the methods available is out of the scope of this article, so we refer the reader to the survey [30] and the book [31].

There is a close relationship between head pose and the distribution of facial landmarks. Huang *et al.* [32] proved that under the weak perspective model the 3D pose of a 3-point configuration is uniquely determined up to a reflection by its projection. Under such circumstance, many researchers use 3D head model points and 2D image projections correspondences to estimate head pose. Hu *et al.* [10] roughly estimated head pose using the asymmetric characteristic of the facial features and refined the result by using a 3D-to-2D model. As the development of face alignment [28], [33], [34], the precision, real-time and robustness have improved, the technology has been wildly used for head pose estimation [11]–[13]. However, since the face model is pre-defined, and there are individual differences in the shape of the participants' face, the model cannot fit the completed face, this results in a certain error in the estimation result. Though deformable models could reduce the error by deforming the head model to adapt to each participant [14], [15], the process of deforming requires a significant amount of data and can be computationally expensive.

Compared with model-based methods, appearance-based methods make full use of image information to estimate the head pose. Appearance-based methods can be divided into classification methods, manifold learning methods and nonlinear regression methods. Head pose estimation is considered as a multiclass classification problem in Classification methods [16]–[18]. Though classification methods are easy to be implemented and have strong real-time performance, they can only predict the approximate range of the head pose

and suffer from the granularity of the estimated angles given the difficulty of training two classes whose angles are very close. So their field of applications are limited.

Manifold embedding methods use feature extraction techniques to create a discriminative feature space for head pose estimation, where the correspondence between the feature space location and the pose is easy to establish. Diaz-Chito *et al.* [19] extracted the initial features based on conventional HOG features [35], and then projected the features onto a feature manifold based on Generalized Discriminative Common Vectors(GDCV). Finally, the head pose is estimated from a continuous regression composed of split fitting and multivariate local regression. Haj *et al.* [21] and Drouard *et al.* [22] reduced the feature dimensions by projecting features to latent space and regressing the head pose from the output of the latent space. Compared with deep learning, manifold embedding methods use fewer computing resources and do not require a large number of training samples. After several years of development, the precision of manifold embedding methods has improved considerably, but its accuracy and robustness still need to be improved for practical applications.

Nonlinear regression methods use a labeled training set to create a nonlinear mapping from images to poses, and CNNs are part of these methods. Because CNNs have the ability to reduce dimensions and extract features automatically, they have achieved good results in various fields. At present, several head pose estimation methods based on CNNs have been proposed. Patacchiola and Cangelosi [9] evaluated the performance of different CNN architectures and different adapter gradient methods on released in-the-wild head pose datasets. Liu *et al.* [23] generated a realistic head pose dataset using rendering techniques and evaluated their CNN-based method on synthetic as well as real data. Ahn *et al.* [24] proposed a multi-task convolutional network for face detection, bounding box refinement and head pose estimation. While the use of CNN has improved the precision of head pose estimation considerably, the excellent performance is only exhibited in the same type of images and conditions present in the training set due to severe overfitting to the training set [19].

It has been demonstrated that there is a close relationship between head pose and the distribution of the landmarks [36], [37]. Ren *et al.* [34] introduced head pose into the processing of face alignment to improve the accuracy of face alignment. Xu and Kakadiaris [38] used global and local CNN features to solve head pose estimation and landmark detection tasks jointly. Similarly, introducing facial landmarks information into the processing of head pose estimation could also improve the accuracy of head pose estimation theoretically. However, this method has not been attempted to date. Given the lack of satisfactory work, we propose a new CNN-based method for head pose estimation according to facial landmarks location and image. Inspired by Deep Alignment Network (DAN) [39] and Boundary-Aware Face Alignment [40], the proposed method uses facial landmarks to normalize face shape into a canonical shape and generate landmark heatmap in the basis of transformed facial landmarks to assist feature extraction. Thanks to the landmark heatmap, the features extracted by CNN correspond better to the pose angles, which will improve the generalization ability of CNN-based methods. Through experiments, we analyze the improvement of accuracy which is caused by introducing information of the facial landmarks into the CNN, and compare the proposed method with state-of-the art head pose estimation methods. The next section will focus on how the proposed method introduces facial landmarks information into the CNN.

III. PROPOSED METHOD

In this section, we firstly give an overview of the proposed method for head pose estimation. Next sections discuss the details of the proposed method.

A. OVERVIEW OF THE METHOD

Fig. 1 shows an overview of the proposed method for head pose estimation. We localize the landmarks based on face alignment techniques. According to the landmarks, a certain area around the face is intercepted and the it is resized to the specified size as the input of CNN. The details of this part are described in subsection B. Our CNN consists of a task simplifier, a heatmap generator and a feed-forward neural network. Referring to Cascade Shape Regression (CSR) framework [33], [34], [41], the proposed method defines a canonical face shape S_0, and normalize the input face shape S_1 into canonical face by affine transformation to simplify the head pose estimation task and improve accuracy. The details of the task simplifier are described in subsection C. Landmark heatmap is created in the basis of the landmarks after affine transformation in the heatmap generator, which are described in subsection D. The warped image is stacked with the landmark heatmap as input for the feed-forward neural network to estimate head pose angles. The structure of the feed-forward neural network is shown in Fig. 2 and described in subsection E. Subsection F details the training procedure.

B. FACE ALIGNMENT

Given the need to use facial landmarks as the input of CNN, we use a state-of-the-art landmark detector (Dlib [28] and FAN [29]) to localize facial landmarks. A certain area around the face in the input image is intercepted based on facial landmarks information, and the face area accounts for a quarter of the total area of the intercepted area, thereby reducing the influence of the scale change of the face image on the estimation accuracy of the pose. The intercepted image and face key information is used as input to our CNN.

C. FACE SHAPE NORMALIZATION

The task simplifier is responsible for aligning input face shape S_1 into a canonical shape S_0 to simplify the estimation task

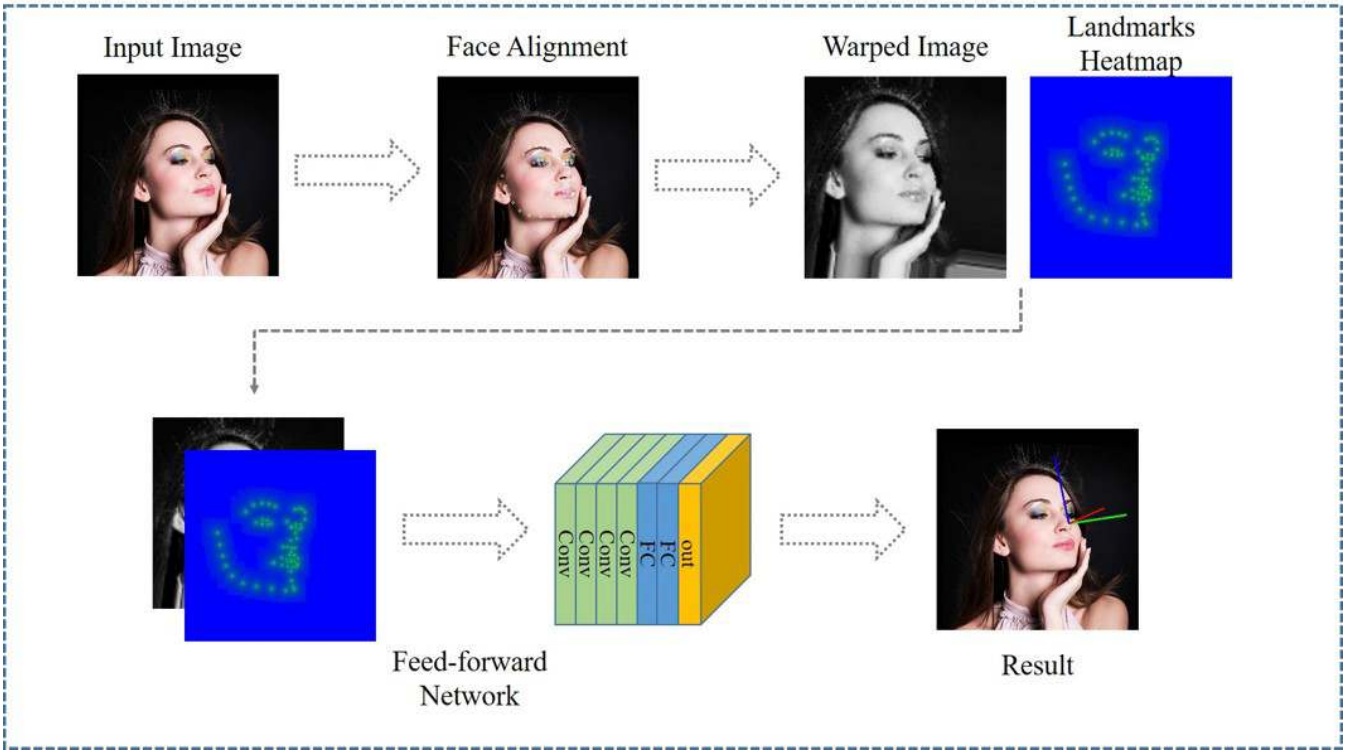


FIGURE 1. Overview of the proposed head pose estimation method.

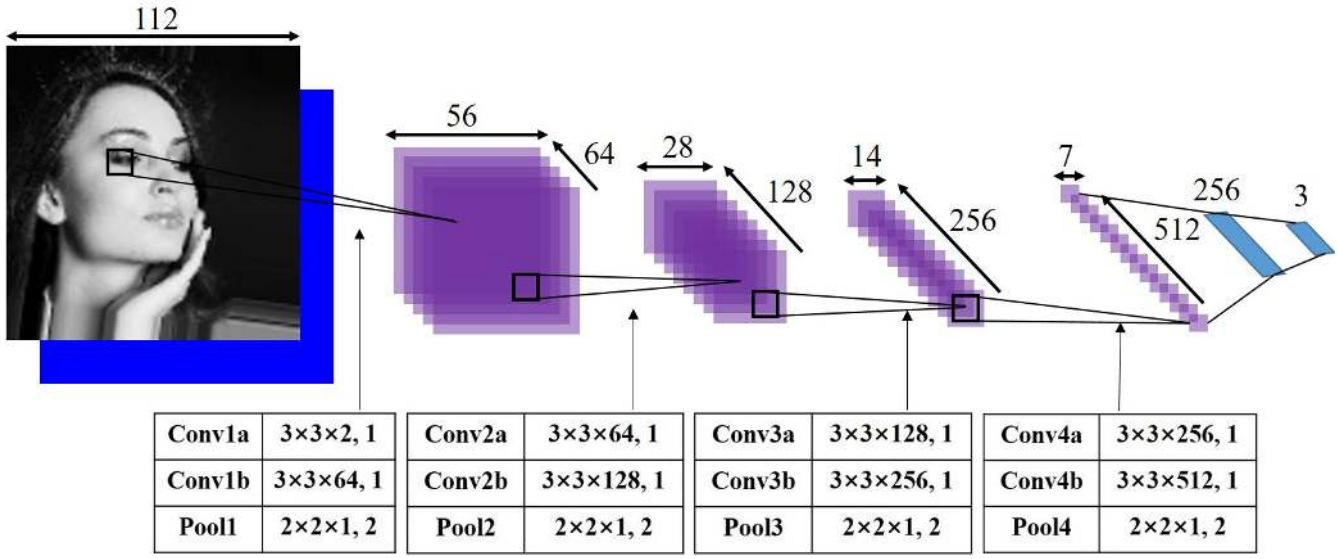


FIGURE 2. Structure of the feed-forward part of our method. The kernels are shown in the figure described as height × width × depth, stride.

and improve accuracy. Referring to Cascade Shape Regression (CSR), the canonical shape consists of n points and the mean of these points is in the center of image, as shown in Fig. 3. The input face shape S_1 and the canonical shape S_0 consist of n points. (x_n, y_n) is the coordinates of the nth point of canonical shape S_0 and (w_n, z_n) is the coordinates of the nth point of inpust shape S_1.

To align the input face shape S_1 to the canonical shape S_0, the translational component and scale component should be removed firstly, as follows:

$$\sigma_0 = \sqrt{\frac{(x_1 - \bar{x})^2 + (y_1 - \bar{y})^2 + \cdots}{n}}$$

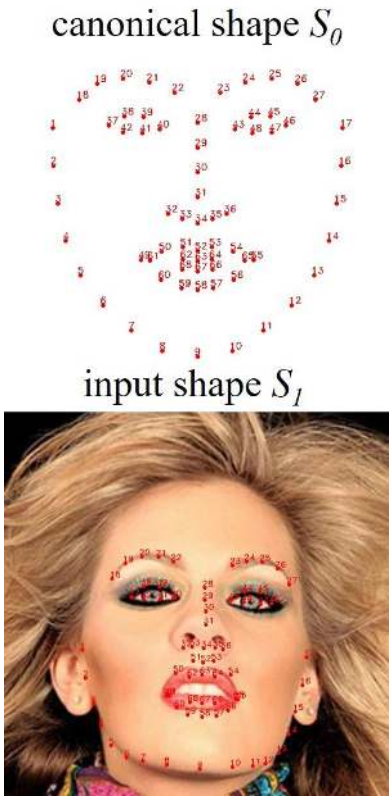


FIGURE 3. Schematic diagram of canonical shape S_0 and input shape S_1.

$$S_0' = \{(\frac{x_1 - \bar{x}}{\sigma_0}, \frac{y_1 - \bar{y}}{\sigma_0}), \cdots\} \tag{1}$$

$$\sigma_1 = \sqrt{\frac{(w_1 - \bar{w})^2 + (z_1 - \bar{z})^2 + \cdots}{n}}$$

$$S_1' = \{(\frac{w_1 - \bar{w}}{\sigma_1}, \frac{z_1 - \bar{z}}{\sigma_1}), \cdots\} \tag{2}$$

$(\bar{x}, \bar{y})$ is the mean of the canonical shape S_0 and $(\bar{w}, \bar{z})$ is the mean of input shape S_1. σ_0 and σ_1 are the root mean square distance of S_0 and S_1, respectively.

Next, calculate the rotation portion of the affine matrix. The rotation portion is an orthogonal matrix $\mathbf{R}$ that the most closely maps S_1 to S_0, and it can be written as follows:

$$\mathbf{R} = \arg \min_{\Omega} \left\| \Omega S_1{}' - S_0{}' \right\|_F^2 \tag{3}$$

Based on the conclusion of Schönemann et al. [42]. $\mathbf{R}$ could be solved as follows:

$$\mathbf{U}\Sigma\mathbf{V}^{\mathbf{T}} = \begin{bmatrix} w_1{}' & \cdots & w_n{}' \\ z_1{}' & \cdots & z_n{}' \end{bmatrix} \begin{bmatrix} x_1{}' & y_1{}' \\ \vdots & \vdots \\ x_n{}' & y_n{}' \end{bmatrix}$$

$$\mathbf{R} = \mathbf{V}\mathbf{U}^{\mathbf{T}} \tag{4}$$

(x_n', y_n') are the points of S_0' and (w_n', z_n') are the points of S_1'. $\mathbf{U}\Sigma\mathbf{V}^{\mathbf{T}}$ is the result of singular value decomposition. The input face shape is scaled after being rotated to the size of canonical face shape S_0. $S_1{}''$ is the input shape after rotation and scaling, and $(\bar{w}'', \bar{z}'')$ is the mean of S_1'', which can be written as follows:

$$\begin{bmatrix} \bar{w}'' \\ \bar{z}'' \end{bmatrix} = \frac{\sigma_0}{\sigma_1} \mathbf{R} \begin{bmatrix} \bar{w} \\ \bar{z} \end{bmatrix} \tag{5}$$

Finally, S_1'' is translated to align the mean of S_1'' with the mean of S_0 and the S_1''' after translation is the most close to S_0. The affine transformation matrix $\mathbf{A}$, which aligns the input face shape S_1 to canonical face shape S_0, is defined as

$$\mathbf{A} = \begin{bmatrix} \frac{\sigma_0}{\sigma_1} r_{11} & \frac{\sigma_0}{\sigma_1} r_{12} & \bar{x} - \bar{w}'' \\ \frac{\sigma_0}{\sigma_1} r_{21} & \frac{\sigma_0}{\sigma_1} r_{22} & \bar{y} - \bar{z}'' \\ 0 & 0 & 1 \end{bmatrix} \tag{6}$$

$r_{11}, r_{12}, r_{21}, r_{22}$ are the parameters of rotation matrix $\mathbf{R}$.

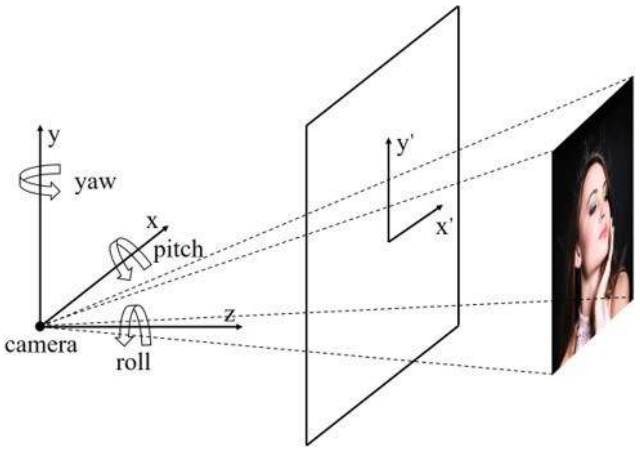


FIGURE 4. Schematic diagram of camera coordinate system.

Fig. 4 shows the schematic diagram of the camera coordinate system. Because of the affine transformation, the pose angles after wrapping are equivalent to the angles which continues to rotate $-\theta$ on the Z axis in the basis of original pose angles. θ can be written as follows:

$$\theta = \tan^{-1}(-\frac{r_{11}}{r_{12}}) \tag{7}$$

The datasets for head pose estimation choose the rotation axes for Euler angle as Z-Y-X. To simplify the calculation, we transform the rotation axes for the Euler angle from Z-Y-X to X-Y-Z. Rotation matrix $\mathbf{M}$ could be calculated from the Euler angles of Z-Y-X. Assuming pitch is α, yaw is β and the roll is γ under X-Y-Z for Euler angles, $\mathbf{M}$ is represented by α, β, γ

$$\mathbf{M} = \begin{bmatrix} \cos(\gamma) & -\sin(\gamma) & 0 \\ \sin(\gamma) & \cos(\gamma) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos(\beta) & 0 & \sin(\beta) \\ 0 & 1 & 0 \\ -\sin(\beta) & 0 & \cos(\beta) \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{bmatrix} \tag{8}$$

Based on eq. 8, α, β, γ can be calculated as follows:

$$\begin{aligned}\alpha &= \tan^{-1}\frac{M_{32}}{M_{33}}\\ \beta &= \tan^{-1}\frac{M_{31}}{\sqrt{M_{11}^2+M_{21}^2}}\\ \gamma &= \tan^{-1}\frac{M_{21}}{M_{11}}\end{aligned} \tag{9}$$

M_{32}, M_{33}, M_{11} and M_{21} are the parameters of rotation matrix $\mathbf{M}$. After affine transformation, the rotation axes for Euler angle is X-Y-Z-Z, which is equivalent to X-Y-Z. At this time, the Euler angles is $(\alpha, \beta, \gamma - \theta)$.

D. LANDMARK HEATMAP

The landmark heatmap is an image where the intensity is highest in the locations of landmarks and it decreases with the distance to the closest landmark. It is first proposed by Kowalski et al. [39] and used in Deep Alignment Networks so that the CNN can focus on the area around the landmarks which are estimated by the previous stage to infer landmark locations. Inspired by this, the heatmap generator is constructed in front of the feed-forward neural network so that the feed-forward neural network could focus on the features around the facial landmark and reduce the error of estimated result caused by background changes. The heatmap is generated by eq.10.

$$H(x,y) = \frac{1}{1+\min_{(w''',z''')\in S_1'''}\|(x,y)-(w''',z''')\|} \tag{10}$$

$H(x, y)$ is the intensity of Point (x, y) of heatmap and S_1''' is the input shape after affine transformation. In the proposed method, the heatmap values are only calculated in a circle of a certain radius around each landmark. Some samples of original images, the results of label, warped images and landmark heatmaps are shown in Fig. 5.

E. FEED-FORWARD NEURAL NETWORK

The overall shape of the feed-forward network is inspired by VGG16 [43]. However, since the head pose estimation task is relatively simple compared to the object detection etc., the input size of other related work is generally small. So we change the input size to 112×112 to improve the real-time of the proposed method. The structure of feed-forward network, kernel size, input shape and output shape are shown in Fig. 2. The feed-forward network consists of eight convolution layers along with two fully connected layers. A max pooling layer with 2×2 region and 2 stride is used to reduce dimension after each two convolution layer. With the exception of max pooling layers and the output layer, every layer takes the advantage of batch normalization and uses Rectified Liner Units (ReLU) for activations to achieve the same accuracy with fewer training steps. A dropout layer is added before the first fully connected layer for regularization to avoid overfitting. The Eular angle of head pose in the range of $[-\pi, +\pi]$, so the output layer uses tanh function for activation and is multiplied by π to normalize the result in the range $[-\pi, +\pi]$.

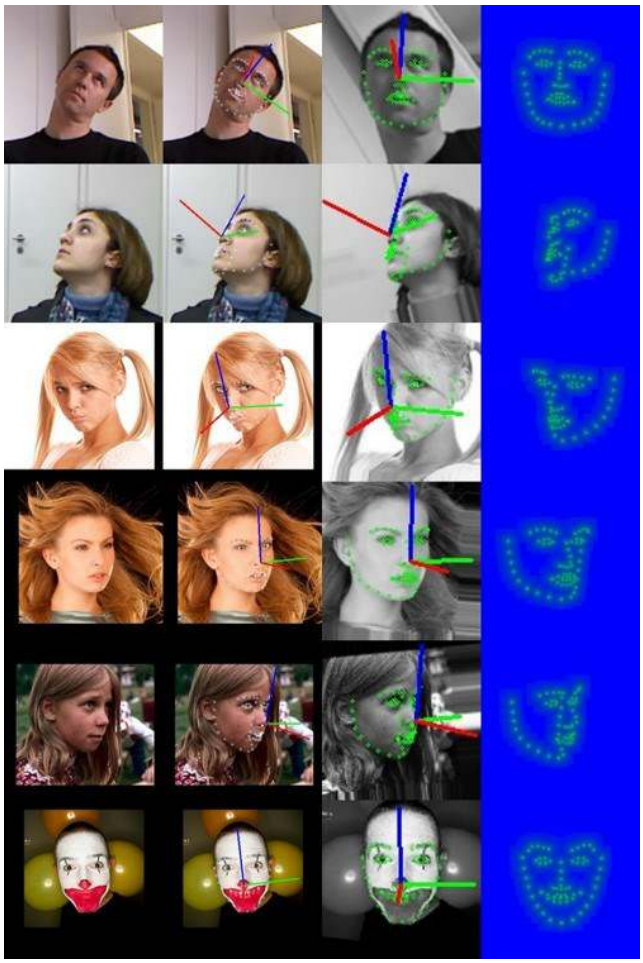


FIGURE 5. Sample images of the dataset relabeled based on AFLW, Biwi head-pose dataset and intermediate results of our proposed method. The columns show: the original images, the results we have relabeled, the intermediate results after face shape normalization and the landmark heatmap.

F. TRAINING PROCEDURE

In the training procedure, the inputs of our method are images, landmark locations and the labeled head pose. The loss function which is used in training is defined as follows:

$$L = \frac{1}{N_b}\sum_{i=1}^{N_b}\frac{\sqrt{\|\mathbf{P}-\mathbf{P}'\|_2^2}}{3} \tag{11}$$

where N_b is the number of mini-batch feeds in training, $\mathbf{P}$ is a vector of labeled head pose after affine transformation and $\mathbf{P}'$ is the estimated pose.

We train our model using adaptive moment estimation (Adam) [44]. The updated rule for Adam can be expressed as follows:

$$w_{t+1} = w_t - \frac{l_r}{\sqrt{m_2}+\varepsilon}m_1 \tag{12}$$

where w_{t+1} and w_t are the weights at time $t+1$ and t. l_r is the learning rate and ε is a small value used to avoid division by zero. The two moments m_1 and m_2 are taken at time t, and before the weights are updated, they are corrected to limit a bias toward zero during the first steps. The moments are regulated by two decaying factors β_1 and β_2. The authors suggest that these parameters be initialized to standard values $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We used these values in our experiment.

IV. EXPERIMENTS AND RESULTS

A. DATASETS AND EXPERIMENTAL SETTING

1) DATASETS

We evaluate the proposed method using six datasets as follows:

- 300W-LP [45]: 300W [46] standardizes multiple alignment datasets with 68 landmarks, including AFW [47], LFPW [48], HELEN [49], IBUG [46] and XMSVTS [50]. With 300W, Zhu *et al.* [45] adapted 3D Dense Face Alignment (3DDFA) to generate 61255 samples across large poses (1786 from IBUG, 5207 from AFW, 16556 from LFPW and 37676 from HELEN. XM2VTS is not used), which is further expanded to 122450 samples using flipping. We train our CNN on 300W-LP.
- AFLW2000-3D [45], [51]: AFLW2000-3D contains the first 2000 identities of the AFLW, which have been re-annotated with 68 3D landmarks using 3DDFA [45]. There are 1306 samples in $[0°, 30°]$, 462 samples in $[30°, 60°]$ and 232 samples in $[60°, 90°]$.
- Biwi Dataset [52]: The dataset contains approximately 15000 images of 20 people (6 females and 14 males −4 people were recorded twice). For each frame, a depth image, the corresponding RGB image and the annotation of face area and head pose are provided. The absolute yaw degrees within $[0°, 75°]$ and the absolute pitch degrees within $[0°, 60°]$. We have added the label of 68 landmarks for each frame with FAN [29] and relabeled the frames that have not been labeled by FAN with manually.
- CASPEAL [53]: The CASPEAL contains 21840 images of 1040 subjects with pose annotation. For each subjects, images across 21 different poses without any other variation are included. The absolute yaw degrees within $[0°, 60°]$ and the absolute pitch degrees within $[0°, 45°]$. We have added the label of 68 landmarks for each frame with FAN[27] and relabeled the frames that failed to be labeled by FAN manually.
- DrivFace [6]: The DrivFace is composed of 606 samples, acquired over different days from 4 driver (2 women and 2 men). The annotation contains the face bounding box, the facial key points (eyes, nose and mouth) and a set of labels assigning each image into 3 possible gaze direction. The absolute yaw degrees are within $[0°, 45°]$.
- Driver Pose: The Driver Pose is a new dataset proposed by us for testing the generalization ability of the result trained on 300W-LP. It is composed of 3426 samples, acquired from 12 drivers (2 women and 10 men) with three cameras from different directions. The annotation contains the face bounding box and a set of labels assigning each image into 12 possible gaze direction. There are 1256 samples in $[0°, 30°]$, 1492 samples in $[30°, 60°]$ and 678 samples in $[60°, 90°]$.

Some cases of these datasets are shown in Fig. 6.

2) EXPERIMENTAL SETTINGS

Without specifications, we implement our experiment in the basis of Python code and TensorFlow framework [54] in Ubuntu 16.04. The hardware configuration is as follows: NVIDIA TITAN Xp graphics card, 12GB GPU memory, i9-7900X @3.60GHz × 10 processor and 32GB RAM.

To enhance the generalization ability of the training result, we expand the training set by panning, rotating, scaling and mirroring randomly 20 times (except 300W-LP, the dataset has been expanded when proposed). The input size is 112×112. The proposed methods in the experiment are trained for 100 epochs using Adam optimization with a learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The decay rate is 0.96 and learning rate will be decayed after each 2 epochs training.

We compute the *mean average error* (MAE) and *accuracy* on validation set to compare the performance among different method. Referring to Sundararajan and Woodard [55], the *accuracy* has been measured by dividing the range $[-90°, +90°]$ into steps of $\pm 15°$, and it is intended as the percentage of images within 15° of error. The *MAE* is defined as Eq. 13. $(\alpha_e, \beta_e, \gamma_e)$ is the estimated Eular angle of head pose and $(\alpha_g, \beta_g, \gamma_g)$ is the labeled Eular angle of head pose. N_v indicates the number of samples in validation set.

$$MAE = \frac{\sum_{i=1}^{N_v} \left(|\alpha_e - \alpha_g| + |\beta_e - \beta_g| + |\gamma_e - \gamma_g| \right)}{3N_v} \tag{13}$$

B. PERFORMANCE IMPROVEMENT BROUGHT ABOUT BY FACIAL LANDMARKS

To investigate the influence of introducing task simplifier and heatmap generator before feed-forward network, we validate our method, our feed-forward network with only heatmap generator, our feed-forward network with only task simplifier and only our feed-forward network on the same dataset.

We perform two experiment, the one is to validate these methods on AFLW2000-3D based on five-fold cross validation and the other one more challenging is to use the entire 300W-LP as the training set and the entire AFLW2000-3D as the validation set. While training, we validate the result on the whole training set and validation set after each 2 epochs training finished and then record the loss. In the basis of the data, we draw the loss convergence curves of training set and validation set among different methods, the curves are shown in Fig. 6 and Fig. 7. The *mean average error* in degrees and the *accuracy* of training result among different methods are shown in table1 and table2.

FIGURE 6. Cases of datasets used in validation. Rows 1, 3, 5, 7, 9 and 10 are the original images from the CASPEAL dataset, AFLW2000-3D dataset and BIWI dataset. Rows 2, 4, 6 and 8 are the landmarks information that was labeled for these datasets. Landmarks information will be used in our method to improve the accuracy of head pose estimation.

Compared with head pose estimation in the basis of only raw image, head pose estimation in the basis of affine image exhibits better performance in the two experiments. On the validation set, the errors of Eular angles were reduced, especially roll angle (by 2.60° in five-fold cross validation on AFLW2000-3D and by 3.31° in trained on 300W-LP). Task simplifier narrows the variation of head poses by rotating the image on Z-axis. Rotating image on Z-axis mainly narrows the variation of roll of head pose, and the contribution of task simplifier for *mean average error* (MAE) is mainly focused on the roll, which shows that narrowing the variation of head pose is beneficial to simplify the learning task and improve the *accuracy* of head pose estimation. In the condition of insufficient training data (five-fold cross

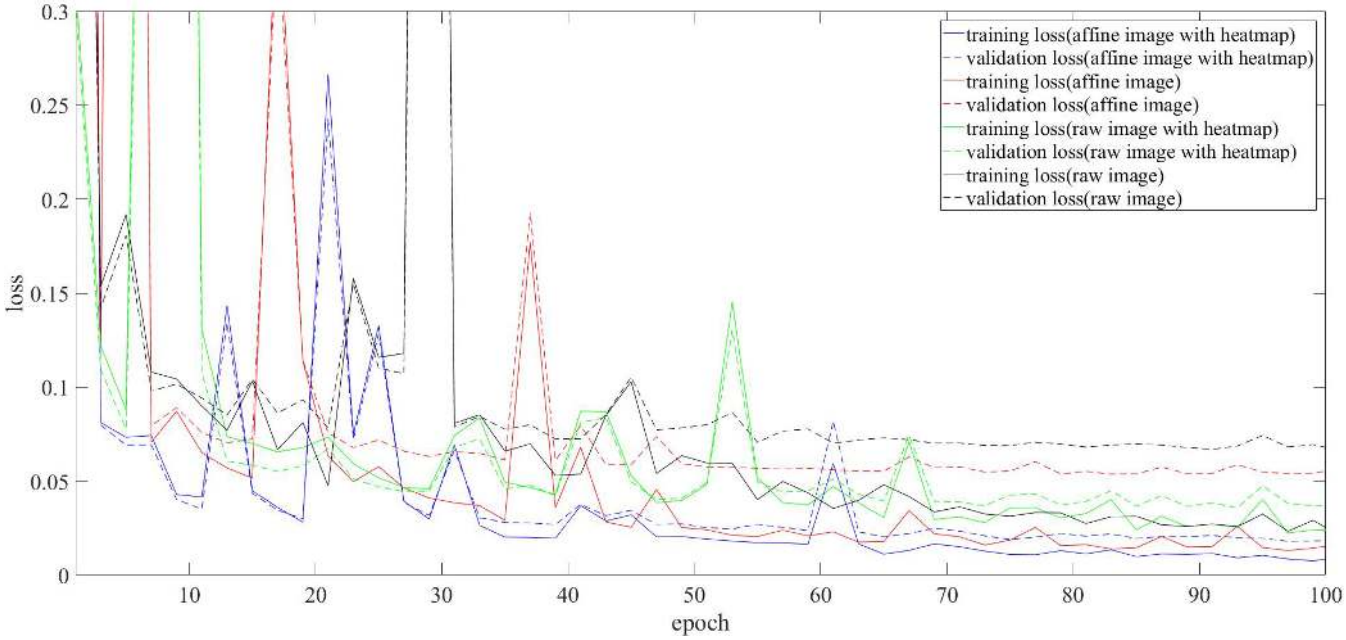


FIGURE 7. Five-fold cross validation on ALFW2000-3D. Solid curves denote training loss, and dotted curves denote validation loss. Blue: head pose estimation based on affine image and heatmap. Red: head pose estimation based on affine image. Green: head pose estimation based on raw image with heatmap. Black: head pose estimation based on raw image.

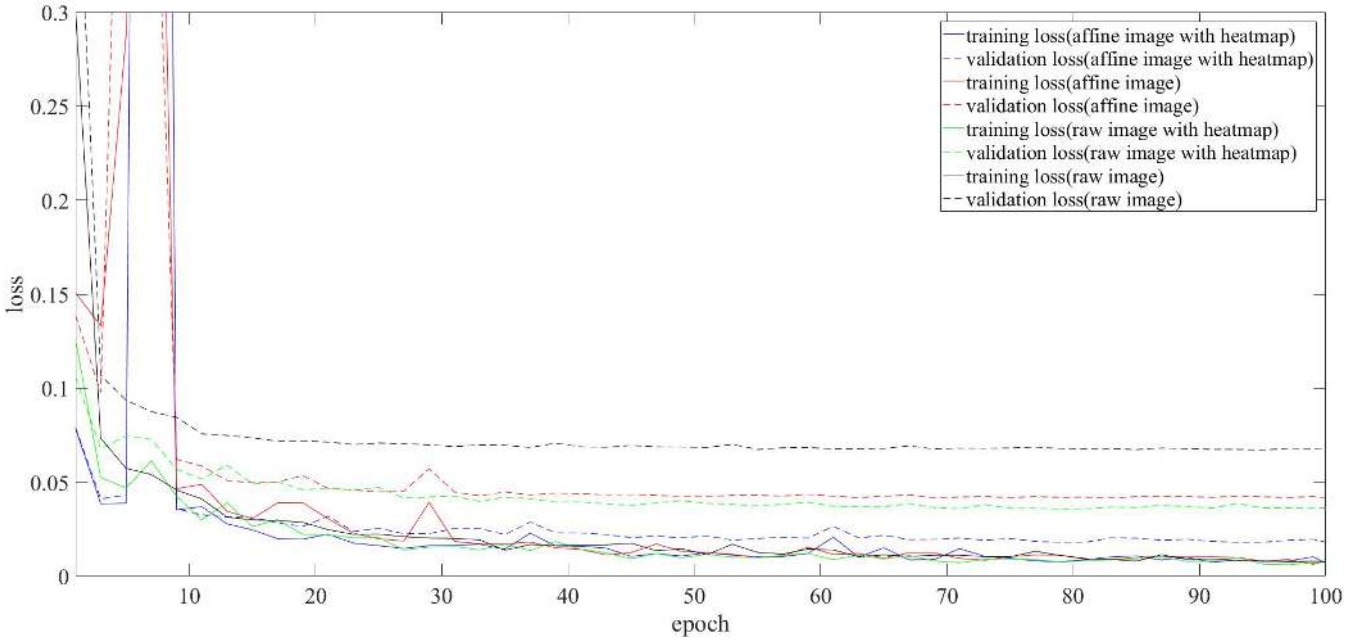


FIGURE 8. Training on 300W-LP and validating on AFLW2000-3D. Solid curves denote training loss, and dotted curves denote validation loss. Blue: head pose estimation based on affine image and heatmap. Red: head pose estimation based on affine image. Green: head pose estimation based on raw image with heatmap. Black: head pose estimation based on raw image.

validation on AFLW2000-3D), the improvement of task simplifier for accuracy is limited (by 4.75%). But in the condition of sufficient training data (trained on 300W-LP and validated on ALFW2000-3D), introducing task simplifier before feed-forward network can improve the accuracy greatly (by 7.60%). This is because the task simplifier does not significantly reduce background interference. The CNN still needs to be trained with a large amount of data to extract the features that have strong correlation with head pose.

Compared with head pose estimation in the basis of only raw image, head pose estimation in the basis of raw image with heatmap can reduce the *mean average error* (MAE) of

TABLE 1. ***Mean Average Error*** **(MAE) in degrees and** ***accuracy*** **(% of images with ±15° error) with different proposed architecture on AFLW2000-3D (five-fold across validation).**

Method	Yaw°	Pitch°	Roll°	MAE°	Accuracy
Ours(affine image with heatmap)	**0.75**	**1.59**	**1.26**	**1.21**	**98.75%**
Ours(heatmap with raw image)	1.14	4.13	4.03	3.98	97.75%
Ours(affine image)	3.74	6.54	3.98	4.75	90.75%
Ours(raw image)	4.24	7.61	6.58	6.14	86.00%

TABLE 2. ***Mean Average Error*** **(MAE) in degrees and** ***accuracy*** **(% of images with ±15° error) with different proposed architecture on AFLW2000-3D (trained on 300W-LP).**

Method	Yaw°	Pitch°	Roll°	MAE°	Accuracy
Ours(affine image with heatmap)	**0.63**	**2.05**	**1.10**	**1.46**	**99.00%**
Ours(heatmap with raw image)	1.25	3.91	3.99	3.05	96.95%
Ours(affine image)	2.51	4.81	3.19	3.50	96.15%
Ours(raw image)	3.99	7.32	6.50	5.94	88.55%

validation set significantly in the two experiments (by 2.16° in five-fold cross validation on AFLW2000-3D and by 2.89° in trained on 300W-LP). In addition, the introduction of heatmap generator can greatly improve the *accuracy* in the both cases of insufficient training data (five-fold cross validation on AFLW2000-3D) and sufficient training data (trained on 300W-LP and validated on ALFW2000-3D) (by 11.75% in the case of insufficient training data and by 8.40% in the case of sufficient training data). Taking advantage of heatmap generator allows obtaining a training result with strong generalization ability in the condition of insufficient training samples. This is due to the introduction of heatmap map, which makes CNN focus more on the area around the facial landmarks when extracting features and reduces the interference of background significantly, the CNN can extract the features that express head pose efficiently more easily.

Task simplifier improves generalization ability by simplifying the learning task and heatmap generator improves generalization ability by making CNN focus on the area around facial landmarks. Since task simplifier and heatmap generator improve generalization ability of training result based on different methods, the lifting effect can be superimposed to a certain extent and the loss of validation set can be reduced further. Introducing both task simplifier and heatmap generator before feed-forward network, the accuracy in the condition of insufficient training samples (five-fold cross validation on AFLW2000-3D) is improved from 86.00% to 98.75% and the accuracy in the condition of sufficient training samples (trained on 300W-LP and validated on ALFW2000-3D) is improved from 88.55% to 99.00%.

C. COMPARISON WITH STATE-OF-THE-ART METHODS

To ensure an exhaustive comparison with other state-of-the-art methods, our method is validated using three publicly available standard datasets, AFLW2000-3D, BIWI, and CASPEAL.

TABLE 3. ***Mean Average Error*** **(MAE) in degrees among different methods on AFLW2000-3D. † trained on 300W-LP.**

Method	Yaw°	Pitch°	Roll°	MAE°
Ours(affine image with heatmap) †	**0.63**	**2.05**	**1.70**	**1.46**
Ours(raw image) †	3.99	7.32	6.50	5.94
Multi-Loss ResNet50 [25](α=1) †	6.92	6.64	5.67	6.41
Multi-Loss ResNet50 [25](α=2) †	6.47	6.60	5.44	6.16
3DDFA [45]	5.40	8.53	8.25	7.393
FAN [29](12 Points)	6.36	12.28	8.71	9.12
Dlib [28](68 Points)	23.15	13.63	10.55	15.77
Ground truth landmarks	5.92	11.76	8.27	8.65

TABLE 4. ***Mean Average Error*** **(MAE) in degrees among different methods on the BIWI dataset. ∗ these methods use depth information.**

Method	Yaw°	Pitch°	Roll°	MAE°
Ours(affine image with heatmap)	2.83	5.52	**2.86**	3.74
Ours(raw image)	**2.39**	4.92	3.09	3.47
Famelli et al. [52] ∗	3.50	3.80	5.40	4.23
Wang et al. [56] ∗	8.8	8.5	7.40	8.23
Drouard et al. [22]	4.24	5.43	4.13	4.60
Liu et al. [23]	6.10	6.00	5.70	5.93
Ruzi et al [25]	3.29	**3.39**	3.00	**3.23**

TABLE 5. **Angular error of pitch and yaw in degrees among different methods on the CASPEAL dataset.**

Method	Yaw°	Pitch°
Ours(affine image with heatmap)	**0.94**	**0.66**
EL_TT_N [57]	4.97	-
Diaz-Chito et al [19]	3.68	13.53
OPENFACE [58]	16.74	24.38
DRMF [59]	11.70	27.68
HPE [22]	15.77	36.29

Firstly, we compare our method with other state-of-the-art methods on AFLW2000-3D. The † indicates that the method is trained on 300W-LP. Multi-Loss ResNet50 [25] estimates head pose angles directly from image intensities based on a CNN. The main task of 3DDFA [45] s to align facial landmarks using a dense 3D model. As a result of the 3D fitting process, a 3D head pose is produced. FAN [29] and Dlib [28] are state-of-the-art landmark detectors and the proposed method is compared with pose estimated from landmarks using the two landmarks detectors and ground truth landmarks. The result is presented in Table 3.

After removing the task simplifier and heatmap generator, our method exhibits performance similar to Multi-Loss Resnet, 3DDFA on AFLW2000-3D. This result is also in line with the conclusion of Patacchiola and Cangelosi [9]: when the number of convolution layers and parameters reach a certain value, adding another convolution layer or more

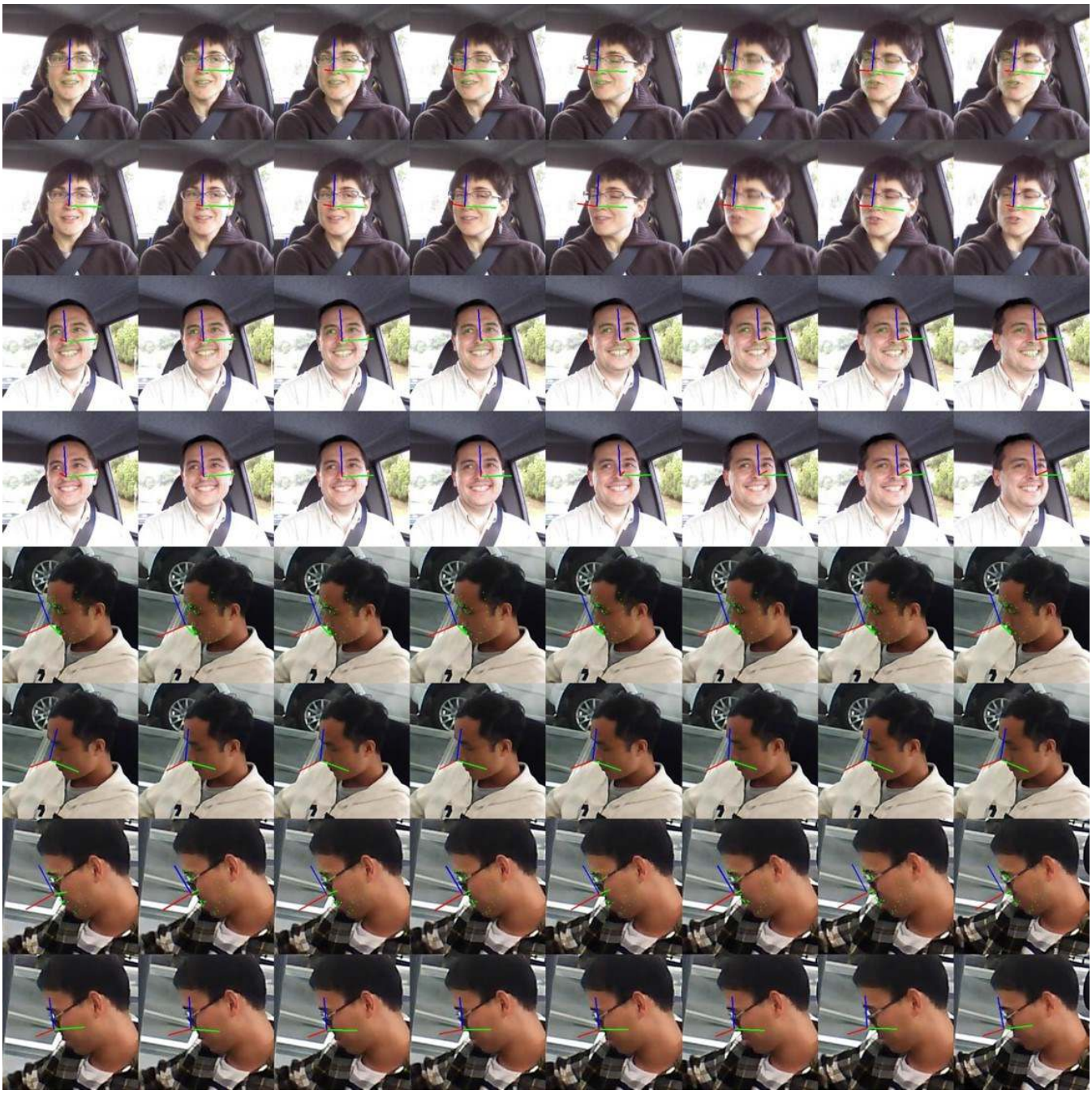


FIGURE 9. Examples of estimation of facial landmarks and head pose in DrivFace and Our dataset. Rows 1, 2, 3, and 4 are the examples of DrivFace dataset and rows 5, 6, 7 and 8 are the examples of our dataset. Rows 1, 3, 5 and 7 show the result estimated by FAN and our method (with landmarks) and rows 2, 4, 6, and 7 show the result estimated by FAN and our method (without lamdmarks).

parameters did not lead to any improvement. By introducing landmarks information into CNN based on the proposed method, the CNN exhibits excellent performance on AFLW2000-3D. This phenomenon further proves our hypothesis: what limits the performance of CNNs on wild head pose dataset is not the expressivity of CNNs, but because the features extracted by the CNN could not reflect head pose efficiently. Because of large variation in head poses, model-based methods that only use landmarks information could not exhibit excellent performance on AFLW2000-3D either. The proposed method improves the *accuracy* and generalization ability of CNN on head pose estimating problem considerably by combining the advantages of both appearance-based methods and model-based methods.

Then, we evaluate the proposed method on the BIWI dataset with five-fold cross-validation at the video level and compare the method with state-of-the-art methods. We also evaluate the proposed method after removing the

task simplifier and heatmap generator on BIWI to analyze the influence of introducing landmarks information into the CNN on the dataset under controllable conditions. The $*$ indicates that depth information is used for head pose estimation in this method. Drouard *et al.* [22] estimated head pose by using the manifold embedding approach and adopted the leave-one-out evaluation protocol at the individual person level. Liu *et al.* [23] and Ruzi *et al.* [25] estimated head pose based on CNN and adopted cross-validation protocol at the video level. The result is presented in Table 4.

Because BIWI is the dataset under controllable conditions (with little background interference), the features extracted by CNN could reflect the head pose more efficiently. In this case, what primarily limits the performance on validation set theoretically is the expressivity of the CNN. The expressivity of the CNN is mainly dependent on its number of layers. In all methods based on CNNs listed in Table 4, the order of expressivity from low to high is Liu *et al.* [23], ours, and then Ruzi *et al.* [25], and the performance on BIWI from low to high is the same as that of expressivity. The experimental results are consistent with our estimate. Thus, introducing landmarks information into CNNs according to the proposed methods can not improve the performance on the datasets under controllable conditions significantly. However, situations under controllable conditions are rare in practice and the results trained on such datasets are less practical. Compared to other methods, all methods based on CNNs exhibit excellent performance and are shown to be promising.

Finally, we evaluate the proposed method on CASPEAL to test the performance when only estimating a single angle. CASPEAL is only annotated with yaw and pitch. we change the output layer from 3 to 1 and evaluate the proposed method on CASPEAL with five-fold validation for twice, which predicts yaw and pitch separately. EL_TT_N [57] estimates the yaw of the head pose based on deformable model and is evaluated on 940 subjects of CASPEAL. Diaz-Chito *et al.* [19] and HPE [22] divide CASPEAL into two parts. Diaz-Chito *et al.* evaluate their method on each part with two-fold cross-validation and repeat 10 times with different random training/testing sample choices. HPE [22] is evaluated on each part with five-fold cross validation. Discriminative Response Map Fitting (DRMF) [59] and OPENFACE [58] are also added to the comparison for reference according to their results, but using the best trained model provided by the authors.

The result is presented in Table 5. Compared with other methods, the proposed method still exhibits excellent performance when estimating a single angle of head pose.

D. VALIDATION ON UNKNOWN DATASET

To further compare the generalization ability of the proposed method and the method after removing the task simplifier and heatmap generator, we evaluate two methods on DrivFace [6] and Driver Pose dataset. The both methods are trained on 300W-LP. The facial landmarks are located by FAN using the best trained model provided by the authors.

Representative examples are shown in Fig. 9. The accuracies of the two methods have both reached 99.67%, because there is no large variation of head pose and lighting conditions on DrivFace. Driver Pose is more challenging because there are more large poses on the dataset and larger variation in the lighting condition. Through Fig. 9, we can intuitively find that after introducing landmarks information into the CNN, the accuracy of the CNN is significantly improved in the case of large poses. By introducing landmarks information into CNN based on the proposed method, the estimation accuracy on Driver Pose has been improved from 87.9% to 99.3%. The experiment shows that the proposed method could improve the generalization ability of trained result considerably.

V. CONCLUSION

This paper presents a new method to introduce landmarks information into a CNN for improving the accuracy of estimation and the generalization ability of the trained result. A task simplifier and a heatmap generator are constructed before the feed-forward neural network. In the task simplifier, the input face shape is normalized to a canonical shape based on landmarks information to simplify the head pose estimation task. A heatmap is generated in the heatmap generator to make the CNN focus on the area around the landmarks while extracting features and the warped image is stacked as the input of feed-forward neural network. The validation result on the wild head pose dataset (300W-LP and AFLW2000-3D) shows that the proposed method can improve the accuracy of head pose estimation in the wild considerably. Comparing with validation result on the datasets under controllable conditions (BIWI and CASPEAL), we find that what limits the performance of CNN on wild head pose datasets is that the features extracted by CNN cannot express the head pose efficiently, given the expressivity of the CNN. After introducing landmarks information into the CNN based on the proposed method, the features extracted by the CNN can reflect the head pose more efficiently. Finally, we test the trained result of 300W-LP on DrivFace and Driver Pose. The result shows that the trained result based on the proposed method has superior generalization capacity. In the next step, we are exploring how to combine the proposed method with the face alignment method into a single neural network to estimate the head pose during face alignment.

REFERENCES

[1] R. Stiefelhagen, "Tracking focus of attention in meetings," in *Proc. 4th IEEE Int. Conf. Multimodal Interfaces*, Oct. 2002, pp. 273–280.

[2] H. Salam, O.Çeliktutan, I. Hupont, H. Gunes, and M. Chetouani, "Fully automatic analysis of engagement and its relationship to personality in human-robot interactions," *IEEE Access*, vol. 5, pp. 705–721, 2017.

[3] R. H. Baxter, M. J. V. Leach, S. S. Mukherjee, and N. M. Robertson, "An adaptive motion model for person tracking with instantaneous head-pose features," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 578–582, May 2015.

[4] B.-G. Lee and W.-Y. Chung, "Driver alertness monitoring using fusion of facial features and bio-signals," *IEEE Sensors J.*, vol. 12, no. 7, pp. 2416–2422, Jul. 2012.

[5] N. Alioua, A. Amine, A. Rogozan, A. Bensrhair, and M. Rziza, "Driver head pose estimation using efficient descriptor fusion," *EURASIP J. Image Video Process.*, vol. 2016, p. 1–14, Jan. 2016.

[6] K. Diaz-Chito, A. Hernández-Sabaté, and A. M. López, "A reduced feature set for driver head pose estimation," *Appl. Soft Comput.*, vol. 45, pp. 98–107, Aug. 2016. doi: 10.1016/j.asoc.2016.04.027.

[7] C. Yin and X. Yang, "Real-time head pose estimation for driver assistance system using low-cost on-board computer," in *Proc. 15th ACM SIGGRAPH Conf. Virtual-Reality Continuum Appl. Ind.*, vol. 1, 2016, pp. 43–46.

[8] G. Borghi, R. Gasparini, R. Vezzani, and R. Cucchiara, "Embedded recurrent network for head pose estimation in car," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2017, pp. 1503–1508.

[9] M. Patacchiola and A. Cangelosi, "Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods," *Pattern Recognit.*, vol. 71, pp. 132–143, Nov. 2017.

[10] Y. Hu, L. Chen, Y. Zhou, and H. Zhang, "Estimating face pose by facial asymmetry and geometry," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2004, pp. 651–656.

[11] P. Martins and J. Batista, "Accurate single view model-based head pose estimation," in *Proc. 8th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2008, pp. 1–6.

[12] A. Dopfer, H.-H. Wang, and C.-C. Wang, "3d active appearance model alignment using intensity and range data," *Robot. Auton. Syst.*, vol. 62, no. 2, pp. 168–176, Feb. 2014. doi: 10.1016/j.robot.2013.11.002.

[13] C. Gou, Y. Wu, F.-Y. Wang, and Q. Ji, "Coupled cascade regression for simultaneous facial landmark detection and head pose estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2906–2910.

[14] M. Krinidis, N. Nikolaidis, and I. Pitas, "3-D head pose estimation in monocular video sequences using deformable surfaces and radial basis functions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, pp. 261–272, Feb. 2009.

[15] X. Yu, J. Huang, S. Zhang, and D. N. Metaxas, "Face landmark fitting via optimized part mixtures and cascaded deformable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2212–2226, Nov. 2016.

[16] B. Han, S. Lee, and H. S. Yang, "Head pose estimation using image abstraction and local directional quaternary patterns for multiclass classification," *Pattern Recognit. Lett.*, vol. 45, pp. 145–153, Aug. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016786551400097X

[17] I. Chamveha *et al.*, "Appearance-based head pose estimation with scene-specific adaptation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 1713–1720.

[18] Z. Zhang, Y. Hu, M. Liu, and T. Huang, "Head pose estimation in seminar room using multi view face detectors," in *Proc. 1st Int. Eval. Workshop Classification Events, Activities Relationships*. Berlin, Germany: Springer, Apr. 2007, pp. 299–304. [Online]. Available: http://dl.acm.org/citation.cfm?id=1759639.1759673

[19] K. Diaz-Chito, J. M. D. Rincón, A. Hernández-Sabaté, and D. Gil, "Continuous head pose estimation using manifold subspace embedding and multivariate regression," *IEEE Access*, vol. 6, pp. 18325–18334, 2018.

[20] D. Huang, M. Storer, F. De la Torre, and H. Bischof, "Supervised local subspace learning for continuous head pose estimation," in *Proc. CVPR*, Jun. 2011, pp. 2921–2928.

[21] M. A. Haj, J. Gonzàlez, and L. S. Davis, "On partial least squares in head pose estimation: How to simultaneously deal with misalignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2602–2609.

[22] V. Drouard, R. Horaud, A. Deleforge, S. Ba, and G. Evangelidis, "Robust head-pose estimation based on partially-latent mixture of linear regressions," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1428–1440, Mar. 2017.

[23] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei, "3D head pose estimation with convolutional neural network trained on synthetic images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1289–1293.

[24] B. Ahn, D.-G. Choi, J. Park, and I. S. Kweon, "Real-time head pose estimation using multi-task deep neural network," *Robot. Auto. Syst.*, vol. 103, pp. 1–12, May 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889017303524

[25] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 2074–2083.

[26] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, Apr. 1980. doi :10.1007/BF00344251.

[27] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[28] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1867–1874.

[29] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem?(And a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1021–1030.

[30] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.

[31] C. Wang, Y. Guo, and X. Song, *Head Pose Estimation Via Manifold Learning*. 2017.

[32] T. S. Huang, A. M. Bruckstein, R. J. Holt, and A. N. Netravali, "Uniqueness of 3D pose under weak perspective: A geometrical proof," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 12, pp. 1220–1221, Dec. 1995.

[33] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 532–539.

[34] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1685–1692.

[35] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.

[36] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson, "Face alignment assisted by head pose estimation," *CoRR*, vol. abs/1507.03148, Jul. 2015. [Online]. Available: https://arxiv.org/abs/1507.03148

[37] S. Zhu, C. Li, C.-C. Loy, and X. Tang, "Unconstrained face alignment via cascaded compositional learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3409–3417.

[38] X. Xu and I. A. Kakadiaris, "Joint head pose estimation and face alignment framework using global and local CNN features," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May/Jun. 2017, pp. 642–649.

[39] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 88–97.

[40] 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.*, Jun. 2018, pp. 2129–2138.

[41] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2887–2894.

[42] P. H. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, Mar. 1966. doi: 10.1007/BF02289451.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, Sep. 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, Dec. 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

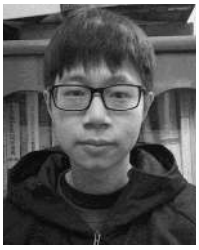
[45] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 146–155.

[46] 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 (ICCVW)*, Dec. 2013, pp. 397–403.

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

[48] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proc. CVPR*, Jun. 2011, pp. 545–552.

[49] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Jun. 2013, pp. 386–391.
[50] 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. Mar. 1999, pp. 965–966.
[51] M. Köstinger, 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*, Nov. 2011, pp. 2144–2151.
[52] G. Fanelli, J. Gall, and L. van Gool, "Real time head pose estimation with random regression forests," in *Proc. CVPR*, Jun. 2011, pp. 617–624.
[53] W. Gao *et al.*, "The CAS-peal large-scale Chinese face database and baseline evaluations," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 1, pp. 149–161, Jan. 2008.
[54] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *CoRR*, vol. abs/1603.04467, Mar. 2016. [Online]. Available: https://arxiv.org/abs/1603.04467
[55] K. Sundararajan and D. L. Woodard, "Head pose estimation in the wild using approximate view manifolds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 50–58.
[56] B. Wang, W. Liang, Y. Wang, and Y. Liang, "Head pose estimation with combined 2D SIFT and 3D HOG features," in *Proc. 7th Int. Conf. Image Graph.*, Jul. 2013, pp. 650–655.
[57] A. Narayanan, R. M. Kaimal, and K. Bijlani, "Estimation of driver head yaw angle using a generic geometric model," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 12, pp. 3446–3460, Dec. 2016.
[58] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.
[59] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3444–3451.



JIAHAO XIA was born in Wuhan, China, in 1995. He received the B.Sc. degree from the Wuhan University of Technology, China, in 2017, and the master's degree from Hunan University, in 2017. His research interests include deep learning algorithm, machine learning, and intelligent vehicle.

LIBO CAO was born in Hengyang, China, in 1964. He received the B.S. degree in vehicle engineering from the University of Hunan, Changsha, China, in 1989, and the Ph.D. degree in mechanical engineering from Hunan University, Changsha, in 2002.

Since 2002, he has been a Doctoral Supervisor with the State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body. He has been a Visiting Scholar with the University of Technology Berlin, Germany, from 2003 to 2004. He had a joint study with Wayne State University, USA. He has authored one book, more than 100 articles, and more than five Chinese national inventions. His research interests include active safety and advance driver-assisted systems, injury biomechanics, passive safety, and self-driving. He is a Reviewer of the *Journal of Automotive Safety and Energy* and *Automotive Technology*.

GUANJUN ZHANG was born in Shandong, China, in 1981. He received the B.E., M.E., and Ph.D. degrees from Hunan University, China, in 2003, 2005, and 2009, respectively. From 2007 to 2008, he was a Visiting Scholar with Wayne State University. Since 2010, he has been with Hunan University. From 2016 to 2018, he was a Visiting Scholar with the University of Michigan. His research interests include image processing, vehicle crashes, automotive restraint systems, and automotive structures.

JIACAI LIAO received the B.E. degree in vehicle engineering from the Hunan University of Mechanical and Vehicle Engineering, Changsha, China, in 2017. He is currently pursuing the Ph.D. degree with Hunan University, Changsha. His research interests include computer vision, machine learning, self-driving, and intelligent transportation systems.