

Location-guided Head Pose Estimation for Fisheye Image

Bing Li, Dong Zhang, Cheng Huang, Yun Xian, Ming Li, *Senior Member, IEEE*, and Dah-Jye Lee, *Senior Member, IEEE*

Abstract—Camera with a fisheye or ultra-wide lens covers a wide field of view that cannot be modeled by the perspective projection. Serious fisheye lens distortion in the peripheral region of the image leads to degraded performance of the existing head pose estimation models trained on undistorted images. This paper presents a new approach for head pose estimation that uses the knowledge of head location in the image to reduce the negative effect of fisheye distortion. We develop an end-to-end convolutional neural network to estimate the head pose with the multi-task learning of head pose and head location. Our proposed network estimates the head pose directly from the fisheye image without the operation of rectification or calibration. We also created a fisheye-distorted version of the three popular head pose estimation datasets, BIWI, 300W-LP, and AFLW2000 for our experiments. Experiments results show that our network remarkably improves the accuracy of head pose estimation compared with other state-of-the-art one-stage and two-stage methods.

Index Terms—Head pose estimation, fisheye camera, fisheye distortion, convolutional neural networks

I. INTRODUCTION

THE goal of head pose estimation (HPE) in the context of computer vision is to estimate the orientation of the head concerning the camera coordinate system. The estimated head pose is usually expressed by Euler angles (pitch, yaw, roll) [1]. Head pose is regarded as an important non-verbal characteristic of human intent for human social behavior analysis [2]. It is involved in many cognitive processes and provides strong cues for human attention, motivation, and intention. Head pose estimation has attracted significant research interest in the fields of cognitive science and computer vision as it helps the computer or robot understand human intention and plays an essential role in various intelligent applications such as driver assistance systems [3] [4], video-conferencing [5], security surveillance [6], virtual reality [7], human-robot interaction [8] and cognitive system [9].

Traditionally, research work on HPE was committed to estimating the head pose from rectilinear (conventional) images. With the rapid growth of the fisheye camera market, estimating the head pose from fisheye image is becoming an emerging requirement.

Bing Li, Dong Zhang, Cheng Huang, and Yun Xian are with the School of Electronics and Information Technology, Sun Yat-sen University, China (e-mail: libing29@mail2.sysu.edu.cn; zhangd@mail.sysu.edu.cn; huangch63@mail2.sysu.edu.cn; xiany7@mail2.sysu.edu.cn).

Ming Li is with the Data Science Research Center, Duke Kunshan University, Kunshan, China, 215316. (e-mail: ming.li369@dukekunshan.edu.cn).

Dah-Jye Lee is with the Department of Electric and Computer Engineering, Brigham Young University, U.S.A. (e-mail: djlee@byu.edu).

Corresponding author: Dong Zhang (e-mail: zhangd@mail.sysu.edu.cn).

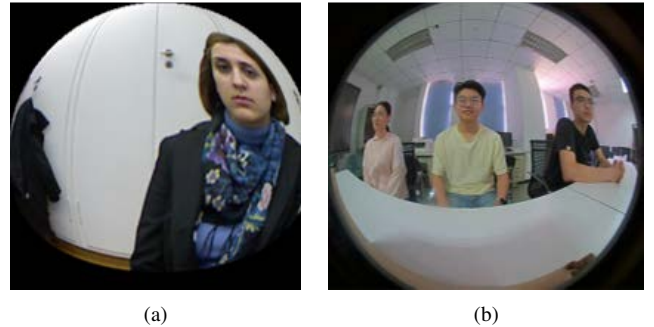


Fig. 1. Examples of fisheye images obtained from (a) BIWI-360 dataset, (b) a fisheye camera (JR®HF900).

Rather than generating images with straight lines of perspective, fisheye camera use a special mapping to produce images with a characteristic convex non-rectilinear appearance, as shown in Fig. 1. This distortion nature of the fisheye camera makes estimating head pose from fisheye image a very challenging task.

Head pose estimation from rectilinear image has already achieved promising results [10]. An intuitive idea for estimating the head pose from fisheye image is the so-called two-stage strategy. It involves rectifying the fisheye image to a rectilinear image first, and then using a good rectilinear-image-based HPE method to estimate the head pose from the rectified image. The main advantage of this strategy is that head pose estimators already developed for the rectilinear image can be directly migrated to rectified fisheye image. If the camera parameters are available, this strategy may achieve good results. However, the camera parameters are not always available in real-world scenarios and the camera calibration is proved to be a tedious and time-consuming process. The lack of the camera parameters prevents the rectification operation from completely eliminating the geometrical difference in appearance between the rectified image and the corresponding rectilinear image, which decreases the accuracy of the two-stage based strategy in head pose estimation.

Another approach is to perform head pose estimation directly on the fisheye images, which we call the one-stage method. One-stage methods do not require an explicit process of image rectification and are expected to result in a unified optimized framework. In image processing, there have been successful attempts to extend traditional techniques designed for rectilinear images to fisheye images or spherical (omnidirectional) images. Hansen et al. mapped a wide-angle image

to a sphere and developed a scale invariant feature transform (SIFT) algorithm on the sphere to match points between wide-angle images [11]. Additionally, Cruz-Mota et al. proposed a SIFT-based algorithm in spherical coordinates for point matching between spherical and rectilinear images [12]. Demonceaux et al. used the geodesic distance metric between pixels in fisheye image to redefine the basic filter kernels [13]. Delibasis proposed an efficient implementation to redefine the Laplacian of Gaussian kernel using the geodesic distance [14]. Georgakopoulos et al. presented a method of human pose recognition for fisheye images by using Convolutional Neural Networks (CNN) enhanced with geodesically corrected Zernike moments [15]. These works employed a one-stage approach without resorting to image rectification and provided us with valuable insights and confidence in the practicability of one-stage approaches for head pose estimation.

In this paper, we address the problem that estimates head pose directly from a single fisheye image without the operation of rectification or calibration. Considering objects appeared at different locations in the camera field of view suffer distinct distortions in the fisheye image, we argue that the accuracy of head pose estimation for fisheye can be improved if the location of the head in the image can be introduced into the estimation network. We present an end-to-end convolutional neural network to estimate the head pose by proposing multi-task learning of head location and head pose. As it does not require rectification, or the knowledge of the camera parameters of the fisheye camera, the proposed method significantly simplifies HPE computation. Considering there are no publicly accessible fisheye image datasets for head pose estimation, we created a fisheye-distorted version of the three popular public benchmark datasets, BIWI, 300W-LP, and AFLW2000, and performed comparison experiments on the new datasets. Experimental results show that our proposed approach achieved higher accuracy compared with two-stage and other one-stage methods. We also constructed a real-world fisheye dataset to evaluate the effectiveness of the proposed method. Experimental results show that our method decreased the average estimation errors for real-world fisheye images compared with the two-stage methods.

The contributions of our work include:

- (1) We propose an end-to-end estimation network to predict head pose directly from a single fisheye image, without the requirement of rectification or calibration. The proposed network estimates the head pose by jointly learning the head pose and head location, and uses the knowledge of head location to reduce the impact of radial distortion for head pose estimation.
- (2) We constructed a fisheye-distorted version of the three popular public benchmark datasets, BIWI, 300W-LP, and AFLW2000, and a real-world fisheye dataset for evaluating the performance of fisheye-image based HPE.
- (3) Experimental results show that the knowledge of the head location in the image contributes to the improved HPE accuracy. The proposed method obtained superior performance compared with two-stage and other one-stage methods on the synthetic datasets of fisheye images.

II. RELATED WORK

Over the last two decades, advances in hardware and software have made perspective (pinhole) cameras used in a wide variety of cognition concerned applications, including human-computer interaction and human-robot interaction. Many efficient methods based on perspective cameras were proposed to capture and estimate the head pose [8], gesture [16], pose of body [17], [18] and gaze direction [19] and achieved promising results. These research works significantly improved the development of intelligent robotic systems and artificial cognitive systems.

With the ability to capture image of a larger scene area than perspective (pinhole) camera, omnidirectional cameras have been widely employed in many applications including video surveillance [20], autonomous driving [21], and 3D reconstruction [22]. Fisheye and catadioptric cameras are two typical omnidirectional camera types. Fisheye camera is a sort of conventional camera combined with a fisheye-shaped lens, while the catadioptric camera combines a standard camera with a shaped mirror. The drawback of catadioptric camera is the shaped mirror is fragile and the camera itself occludes the central part of the image, which makes catadioptric camera a less popular option than the fisheye camera.

Fisheye camera is able to capture images with an ultra-wide field-of-view (FOV). As high-resolution and portable fisheye cameras are becoming relatively affordable, more and more computer vision tasks, such as object detection [20], semantic segmentation [21], fall detection [22], saliency prediction [23], face detection [24], and face recognition [25], have gradually shifted from traditional rectilinear images to fisheye images. However, all existing RGB image-based HPE algorithms were designed for rectilinear images.

Current RGB image-based HPE methods for rectilinear images are commonly grouped into landmark-based methods and landmark-free methods. Landmark-based methods estimate head pose by investigating the geometric relations implied among facial landmarks, while landmark-free methods explore the whole face image and estimate the head pose directly from the input image.

Landmark-based methods detect the facial landmarks and estimate the 3D pose of an object from n 3D-to-2D point-correspondences with Perspective- n -Point (PnP) algorithms. Both classical [26], [27] and deep learning-based [28], [29] facial landmark detectors have been proposed and remarkably improved the accuracy of landmark-based head pose estimation. However, the accuracy of these methods is highly dependent on the correct prediction of the landmark positions. Hence, inferior landmark localization caused by occlusion and extreme rotation can consequently impair the accuracy of head pose estimation.

Rather than exploring the geometric relations among facial landmarks, landmark-free methods estimate the head pose by learning the characteristics of the whole image. Benefitting from the advancement of machine learning techniques, especially the success of deep learning [30] [31] [32], landmark-free methods have achieved promising results on many widely used datasets and become a popular approach for head pose

estimation. Hopenet employed a multi-loss ResNet50 network to predict Euler angles and achieved accurate results by training with both classification and regression losses [33]. FSA-Net proposed a feature aggregation method to improve the accuracy of head pose estimation [34]. Li et al. proposed a lightweight network to obtain highly accurate head pose estimation [35]. They preprocessed the image with image rectification to reduce the negative effect of perspective distortion and employed a discriminative weighted loss to train the network. QuatNet proposed a Quaternion-based head pose regression framework and obtained a more effective HPE method [36]. Img2pose estimated the head pose by calculating six degrees of freedom (6DoF) and 3D face poses of the head directly from raw images without face detection or landmark localization [37].

All the aforementioned HPE methods assume an ideal pinhole camera model is employed. The pinhole camera model mimics the geometrical projection carried out by a pinhole camera and has a small FOV. However, wide FOV cameras violate the rules of perspective transformation and linear projection, and result in distortions in wide-angle image. Therefore, the performance of pinhole camera model degrades significantly for wide-angle image. In order to characterize such distortion, Geyer et al. presented a unifying theory for all central catadioptric systems [38]. A central catadioptric camera means the catadioptric camera equipped with hyperbolic, parabolic, or elliptical mirror. They proved that all central catadioptric systems can be modeled with a projection from a sphere to a plane where the projection center is on the sphere diameter and the plane is perpendicular to the sphere diameter. Furthermore, Ying et al. presented a unified imaging model for fisheye and central catadioptric cameras [39]. With this unified model, a fisheye image can be transformed into a central catadioptric one, and vice versa.

Radial distortion is the main distortion in fisheye camera, which arises from the spherical shape of the lenses and causes the projection onto the image plane to be displaced from the ideal position along the radial axis [40]. As shown in Fig. 1, while the fisheye lens provides ultra-wide view, the face in the image is significantly distorted by the radial distortion of the lens. The radial distortion not only influences the appearance of the face, but also greatly degrades the performance of the existing HPE methods. Hopenet [33] is one of the best HPE models, which was trained on the 300W-LP dataset. We used their pre-trained model and tested it on the fisheye-distorted version of the BIWI-360 dataset we created. The testing results are shown in Fig. 2, where the horizontal coordinate means the normalized radial distance, which represents the distance of the face from the optical axis. The vertical coordinate depicts the average error of the estimated Euler angles in degrees.

Fig. 2 shows the head pose estimation performance of Hopenet [33] is quite good if the location of the head is closed to the center of the fisheye image, but it degrades as head moves toward the perimeter where the distortion is worse. The radial distortion introduced by the fisheye lens is closely related to the distance between the object and the optical axis. When the head is very close to the optical axis or the center of the fisheye image, the accuracy of the current HPE methods is

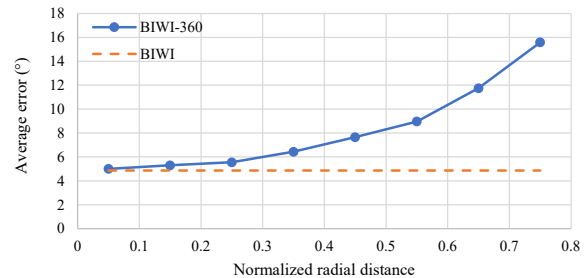


Fig. 2. Average head pose estimation error in fisheye image versus normalized radial distance. The solid line is the result on the fisheye-distorted BIWI-360 dataset. The dotted line is the result on the non-distorted BIWI dataset.

quite good because the radial distortion is minimal. When the head is positioned away from the optical axis of the camera and the radial distortion is more evident, their performance suffers.

Literature review shows that learning related tasks at the same time can achieve better performance than training tasks separately [41]. Studies show that head pose estimation can also achieve better performance with the help of multi-task learning [42], [43]. Ranjan presented an all-in-one convolutional neural network (CNN) for simultaneous face detection, face alignment, pose estimation, gender recognition, smile detection, age estimation, and face recognition [42]. HyperFace first learns common features by CNN, and then simultaneously performs several tasks in a single framework: face detection, landmark localization, gender classification, and head pose estimation [43]. Both of these research work proposed multi-task learning networks that combined the task of head pose estimation with other facial analysis tasks and obtained improved performance for the combined tasks.

Although rapid growth of fisheye camera market has been witnessed in recent years, there is no reported research of HPE specially designed for fisheye image. One compromising approach to estimate head pose from fisheye image is to perform the HPE methods designed for rectilinear image directly on the fisheye image. Without a specially designed mechanism to deal with the distortion existed in fisheye image, the HPE methods for rectilinear image usually can only obtain very limited performance for fisheye image, as shown in Fig. 2. Another approach to estimate head pose from fisheye image is the so-called two-stage strategy which rectifies the fisheye image to a rectilinear image first, and then employs a good rectilinear-image-based HPE method to estimate the head pose from the rectified image. The estimation accuracy of this approach much depends on the performance of rectification methods. As the camera parameters are not always available in real-world applications, the performance of rectification methods is much limited. This situation makes the two-stage strategy also a compromising approach of HPE for fisheye images.

To fill the gap between the potential requirement of HPE for fisheye image and the previous techniques, we propose an end-to-end network to estimate the head pose for fisheye image without the operation of calibration or rectification. We believe

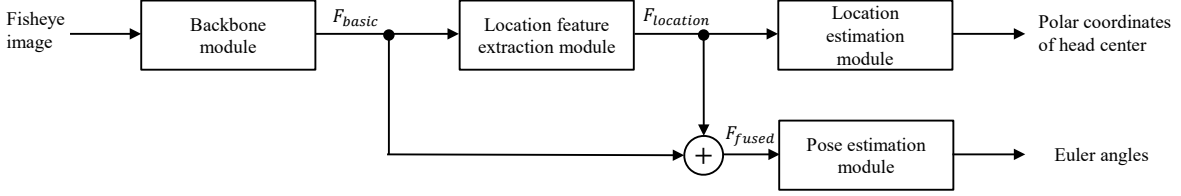


Fig. 3. The overview of the proposed network.

that the accuracy of HPE for fisheye images can be improved if the location of the head in the image can be introduced into the estimation network. The proposed network estimates the head pose by jointly learning the head pose and head location, and uses the knowledge of head location to reduce the impact of radial distortion for head pose estimation. Experimental results show that our approach of using the knowledge of head location achieved higher accuracy compared to two-stage and other one-stage methods.

III. METHODS

A. Network Architecture

The main idea of this work is to use the head location information to help the backbone module learn the camera distortion and improve the HPE accuracy. In this paper, we propose an end-to-end convolutional neural network to estimate the head pose from the fisheye image directly. The architecture of the proposed network, as shown in Fig. 3, consists of the backbone module, location feature extraction module, pose estimation module, and location estimation module.

In the proposed network, the backbone module is first used to extract features from the fisheye image. The obtained features are then fed to the location feature extraction module which learns the location-related features supervised by the ground truth of head location. We fuse the features learned from the backbone module and the location-related features to obtain the fused features. The fused features are fed to an estimation module to estimate the Euler angles of the head pose.

B. Backbone Module

The backbone module is used to extract high-level features directly from the fisheye facial image. Considering the residual neural network (ResNet) [44] is an excellent convolutional neural network that can solve the vanishing gradient problem in deep learning models, and is used in many promising HPE networks for feature learning [33], [43], [45]–[47], we adopt the ResNet50 structure as the backbone of our model. The employed ResNet50 is pre-trained on ImageNet-1k [48] for the object recognition task. The input of the whole network is a single RGB image in size of 224×224. The size of the output feature maps ($\mathbf{F}_{\text{basic}}$ in Fig. 3) from this module is 7×7×2048.

C. Location Feature Extraction Module

As the fisheye distortion of an object presented in the fisheye image is closely related to its location, estimating the

location of the object can help the network learn the cues of distortion. We employ a location feature extraction module to learn the features of head location from $\mathbf{F}_{\text{basic}}$. Specifically, the location feature extraction module generates the location-related features and provides high-level clues implying object distortion. We fuse the location-related features with the basic features and provide location guidance to the estimation of head pose.

To make the network focus on more important features for the task of head location estimation, as shown in Fig. 4, we apply an attention mechanism in the location feature extraction module, which includes a channel attention submodule and a spatial attention submodule successively [49].

For the input feature $\mathbf{F}_{\text{basic}} \in \mathbb{R}^{C \times H \times W}$, the channel attention submodule first generates two feature vectors by using both average-pooling and max-pooling operations, denoted as $\mathbf{F}_{\text{avg}}^C \in \mathbb{R}^C$ and $\mathbf{F}_{\text{max}}^C \in \mathbb{R}^C$.

$$\mathbf{F}_{\text{avg}}^C = \frac{1}{H \times W} \sum_{h=0}^H \sum_{w=0}^W \mathbf{F}_{\text{basic}}(c, h, w), \quad (1)$$

$$\mathbf{F}_{\text{max}}^C = \max_{h,w} \mathbf{F}_{\text{basic}}(c, h, w) \quad (2)$$

After the pooling operations, the two feature vectors are fed into a multi-layer perceptron (MLP) with one hidden layer to produce two weight vectors. The weight-matrices of the hidden layer and the output layer are denoted as $\mathbf{W}_1 \in \mathbb{R}^{(C/r) \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C \times (C/r)}$, where r is the reduction ratio and set to 16. The hidden layer is followed by a ReLU activation function for nonlinearity. Then the two weight vectors are added and activated by a sigmoid function to obtain the channel attention map $\mathbf{M}_C \in \mathbb{R}^C$,

$$\mathbf{M}_C = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{F}_{\text{avg}}^C) + \mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{F}_{\text{max}}^C)) \quad (3)$$

where $\sigma(\cdot)$ and $\delta(\cdot)$ indicate the sigmoid and ReLU function, respectively. Finally, the channel-weighted feature, $\mathbf{F}' \in \mathbb{R}^{C \times H \times W}$, is obtained by multiplying the input feature and the channel attention map.

The spatial attention submodule performs the average-pooling and max-pooling operations along the channel-weighted axis, and generates two feature maps denoted as $\mathbf{F}_{\text{avg}}^S \in \mathbb{R}^{H \times W}$ and $\mathbf{F}_{\text{max}}^S \in \mathbb{R}^{H \times W}$.

$$\mathbf{F}_{\text{avg}}^S(h, w) = \frac{1}{C} \sum_{c=1}^C \mathbf{F}'(c, h, w) \quad (4)$$

$$\mathbf{F}_{\text{max}}^S(h, w) = \max_c \mathbf{F}'(c, h, w) \quad (5)$$

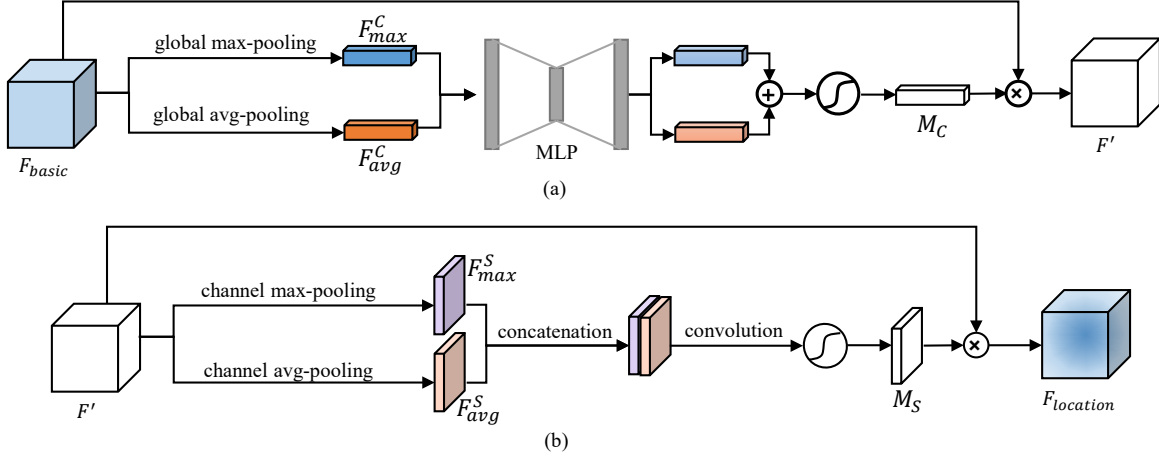


Fig. 4. The location feature extraction module based on the attention mechanism in [49] that includes two sequential submodules: (a) submodule of channel attention and (b) submodule of spatial attention.

After the pooling operations, the two maps are concatenated along the channel axis, and then convolved by a 7×7 convolution layer, producing the spatial attention map $\mathbf{M}_S \in \mathbb{R}^{H \times W}$,

$$\mathbf{M}_S = \sigma(f^{7 \times 7}([\mathbf{F}_{avg}^S; \mathbf{F}_{max}^S])) \quad (6)$$

where $f^{7 \times 7}(\cdot)$ a convolution operation with a 7×7 kernel. Finally, the channel-weighted feature, \mathbf{F}' , is multiplied by the spatial attention map \mathbf{M}_S to produce $\mathbf{F}_{location}$.

D. Location Estimation Module

The location estimation module is used to generate the final estimation of head location in the image and help the module of location feature extraction to learn the cues of location-related distortion. We express the location of the head center with the polar angle and normalized radial distance in polar coordinates. We split the range of polar angle, $\theta \in [-180^\circ, 180^\circ]$ into 72 equal intervals, and the normalized radial distance, $\rho \in [0, 0.99]$ into 66 equal intervals. This design turns head location estimation into a task of classification.

The location estimation module receives $\mathbf{F}_{location}$ from the location feature extraction module as an input feature. A global average pooling layer is first applied to the feature of $\mathbf{F}_{location}$ and produces a one-dimensional (1D) feature vector for two parallel branches to estimate the values of θ and ρ . Each branch includes a fully connected (FC) layer, a softmax layer, and an expectation operation. The softmax layer turns the output of the FC layer into probabilities of θ or ρ in the corresponding intervals. The expectation operation generates the final estimation of θ and ρ by using Eqs. (7) and (8), respectively,

$$\hat{\theta} = 5 \sum_{i=1}^M p_i^\theta \left(i - \frac{M+1}{2} \right), \quad (7)$$

$$\hat{\rho} = 0.015 \sum_{i=1}^N p_i^\rho \left(i - \frac{1}{2} \right), \quad (8)$$

where p_i^θ and p_i^ρ are the probabilities of the estimated polar angle and normalized radial distance in the corresponding i^{th} interval. In Eq. (7), 5 is the interval length of polar angle and M is the number of intervals for the polar angle. In Eq. (8), 0.015 is the interval length of normalized radial distance and N is the number of intervals for the normalized radial distance. The subtracted terms in both equations shift the interval index to the midpoint of the interval.

Although we estimate the values of polar angle and normalized radial distance in the training stage, we do not need to calculate them in the inference stage. We employ the location estimation module to help the location feature extraction module to learn the cues of location-related distortion in the fisheye image. We add the feature of $\mathbf{F}_{location}$ and \mathbf{F}_{basic} to form a fused feature \mathbf{F}_{fused} for the HPE and anticipate the feature of head location to help improve the performance of head pose estimation.

E. Pose Estimation Module

The pose estimation module is used to generate the final estimation of head pose, which is expressed by Euler angles (pitch, yaw, roll). To facilitate the comparison with [33]–[35], we only considered samples whose Euler angles were within the range of $[-99^\circ, 99^\circ]$. The range of each Euler angle is split into 66 equal intervals of 3 degrees. This design turns head pose estimation into a task of classification.

As shown in Fig. 5, a global average pooling layer is first applied to the feature of \mathbf{F}_{fused} and produces a one-dimensional (1D) feature vector for three parallel branches to estimate the values of pitch, yaw, and roll. Each branch employs an FC layer followed by a softmax function. The output of the softmax function is a 1D vector of 66 values that represent the probabilities that the Euler angle of head pose falls into the intervals. Then an expectation layer is used to generate the final estimation of each Euler angle by using

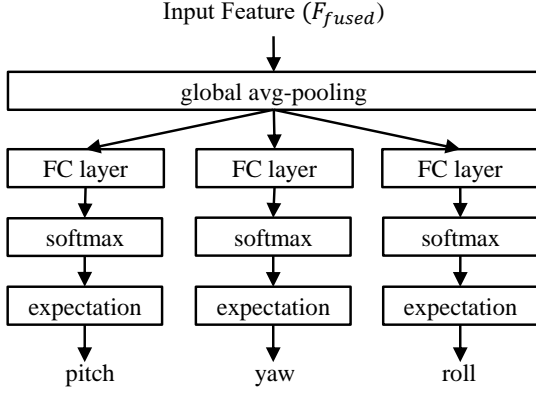


Fig. 5. The structure of the pose estimation module.

Eqs. (9) to (11).

$$\widehat{pitch} = 3 \sum_{i=1}^N p_i^p \left(i - \frac{N+1}{2} \right), \quad (9)$$

$$\widehat{yaw} = 3 \sum_{i=1}^N p_i^y \left(i - \frac{N+1}{2} \right), \quad (10)$$

$$\widehat{roll} = 3 \sum_{i=1}^N p_i^r \left(i - \frac{N+1}{2} \right), \quad (11)$$

where p_i^p , p_i^y , and p_i^r are the probabilities of the angles of pitch, yaw, and roll in the i^{th} interval, respectively, the value 3 before the symbol of summation represents 3 degrees for each interval and N is the number of intervals for each Euler angle ($N=66$ in our design). The subtracted term shifts the interval index to the midpoint of the interval.

F. Loss Function

In our proposed multi-task learning network for pitch, yaw, roll, normalized radial distance and polar angle, the loss function for each task can be formulated by Eq. (12):

$$\mathcal{L}_{task} = \text{CE}(y_{task}, \hat{y}_{task}) + \alpha_{task} \cdot \text{MSE}(y_{task}, \hat{y}_{task}), \quad (12)$$

where the subscript $task$ indicates the estimation task which can be $pitch$, yaw , $roll$, normalized radial distance ρ and polar angle θ ; $\text{CE}(\cdot)$ and $\text{MSE}(\cdot)$ represent cross entropy and mean squared error loss functions, respectively. y_{task} is the ground truth of the corresponding task, while \hat{y}_{task} is the estimated result of the corresponding task, and α_{task} is the weight of $\text{MSE}(\cdot)$ for the corresponding task. We express the location of the head center with the polar angle and normalized radial distance in polar coordinates. The pitch, yaw, roll, and polar angle are estimated in degree, while the estimated value of radial distance is normalized into the range of $[0, 0.99]$ and the range of polar angle is $[-180^\circ, 180^\circ]$. Experiments show the MSE of the estimated normalized radial distance differs from the MSEs of other estimated tasks by at least two orders of magnitude. Therefore, we set the value of α_{task} for the MSE loss of estimated normalized radial distance to 100, while the

values of α_{task} for the MSE losses of estimated pitch, yaw, roll, and polar angle to 1.

The overall loss is represented by a weighted sum of the loss for each task, as shown in Eq. (13):

$$\mathcal{L}_{total} = \mathcal{L}_{pitch} + \mathcal{L}_{yaw} + \mathcal{L}_{roll} + \lambda_1 \mathcal{L}_\rho + \lambda_2 \mathcal{L}_\theta, \quad (13)$$

where \mathcal{L}_{pitch} , \mathcal{L}_{yaw} , \mathcal{L}_{roll} , \mathcal{L}_ρ and \mathcal{L}_θ are the losses for pitch, yaw, roll, normalized radial distance ρ and polar angle θ , respectively; λ_1 and λ_2 are the weights to balance the losses for normalized radial distance and polar angle, respectively.

We provide a list of parameters used to train the network in Table I. The training process of one iteration is summarized in Algorithm 1 for better understanding.

TABLE I
LIST OF PARAMETERS USED TO TRAIN THE NETWORK

Parameter	Note
α_{pitch}	the weight of MSE for the estimation of pitch
α_{yaw}	the weight of MSE for the estimation of yaw
α_{roll}	the weight of MSE for the estimation of roll
α_ρ	the weight of MSE for the estimation of ρ
α_θ	the weight of MSE for the estimation of θ
λ_1	the weight to balance the loss for the estimation of ρ
λ_2	the weight to balance the loss for the estimation of θ

Algorithm 1: Training Process in One Iteration

Input: Fisheye image (I_{input}), model parameters (δ), learning rate (η), backbone module (H_{BAC}), location feature extraction module (H_{LFE}), location estimation module (H_{LOE}), pose estimation module (H_{POE}).

Output: Updated model parameters δ^* ;

- 1: Extract features from fisheye image, $F_{basic} = H_{BAC}(I_{input})$;
- 2: Generate the location-related features, $F_{location} = H_{LFE}(F_{basic})$;
- 3: Generate the fused features, $F_{fused} = F_{location} + F_{basic}$;
- 4: Predict the polar coordinates, $[\hat{\rho}, \hat{\theta}] = H_{LOE}(F_{location})$;
- 5: Predict the Euler angles, $[\hat{pitch}, \hat{yaw}, \hat{roll}] = H_{POE}(F_{fused})$;
- 6: Calculate the loss \mathcal{L}_{total} via Eq. (13);
- 7: Update model parameters, $\delta^* \leftarrow \delta - \eta \frac{\partial}{\partial \delta} \mathcal{L}_{total}$.

IV. EXPERIMENTS AND DISCUSSION

As there is no publicly accessible benchmark dataset of fisheye images for head pose estimation, we created the fisheye-distorted version of three popular public benchmark datasets for our experiments. To demonstrate the superiority of our proposed method, we performed several comparative experiments

with both two-stage and one-stage HPE approaches for fisheye image. Firstly, we compared the estimation accuracy of the proposed network with three two-stage approaches which rectify the fisheye image using a recently developed rectification method and perform head pose estimation using three state-of-the-art methods [33]–[35] trained on the rectilinear datasets. These state-of-the-art methods were included for comparison because their goals are also to predict Euler angles from a single RGB image. Secondly, we compared the estimation accuracy of the proposed network with three existing state-of-the-art (one-stage) methods [33]–[35]. For a fair comparison, these networks were retrained on the fisheye-distorted version of three popular head pose estimation datasets we created. Thirdly, we evaluated the processing speed and number of parameters of the proposed network. The final experiment involved a series of ablation tests to investigate how head location guidance affects the performance of the proposed framework.

A. Datasets

We created datasets of fisheye images annotated with the ground truth of head pose to evaluate the performance of our proposed method. Using a fisheye camera to capture a large number of face images with their accompanying head pose ground truth is a daunting task. Instead, we selected three open-source popular head pose datasets, including BIWI [50], 300W-LP [51], and AFLW2000 [51], and created their fisheye-distorted version. We named these new datasets BIWI-360, 300W-LP-360, and AFLW2000-360, respectively. We followed the same mapping function used in [52]–[55] to transform the original images in the datasets into fisheye-distorted images while keeping the ground truth of the head pose unchanged.

Source Datasets. The BIWI dataset was collected in a lab environment. It contains 15,678 images collected by a Kinect v2 device [50]. Subjects (6 females and 14 males) moved and turned their heads to all possible angles observed from a frontal position. RGB images and depth information were included for each image and head pose annotations were created from the depth information. The head poses of these images are in the range of $\pm 75^\circ$ for yaw, $\pm 60^\circ$ for pitch, and $\pm 50^\circ$ for roll.

Both the 300W-LP [51] and AFLW2000 [51] datasets were created by cropping out the face region directly from photos or video frames downloaded from the Internet or other similar sources. Besides, both datasets use 3D Dense Face Alignment (3DDFA) to fit a morphable 3D model to 2D input images and provide accurate head poses as ground truth. 300W-LP additionally generates synthetic views, greatly expanding the number of images. More specifically, the 300W-LP dataset employed face profiling with a 3D image model to expand the 300W dataset which combines several face alignment datasets including LFPW [56], AFW [57], HELEN [58], IBUG [59], and XMSVTS [60], with 68 landmarks for each face image. As a result, 61,225 samples (16556 from LFPW [56], 5,207 from AFW [57], 37676 from HELEN [58], and 1,786 from IBUG [59]) were generated. 300WP-LP further flipped

all these samples to double the sample size to 122,450. AFLW2000 is a very challenging dataset with large pose variations with various facial expressions and illumination conditions. AFLW2000 provides face images, corresponding head pose ground truth, and 3D face for the first 2000 images in AFLW [61].

Mapping Function. To visually approximate the distortion of a fisheye lens, we first normalized the coordinates of rectilinear image, which means that (0.0, 0.0) represents the origin of the coordinate system located at the image center, and $(\pm 1.0, \pm 1.0)$ represents the coordinates of the four quadrants. Let (x, y) be the normalized coordinates of an arbitrary pixel in a rectilinear image. Its corresponding pixel (x', y') in the fisheye image can be formulated as Eq. (14).

$$(x', y') = \left(x \sqrt{1 - \frac{y^2}{2}}, y \sqrt{1 - \frac{x^2}{2}} \right). \quad (14)$$

Furthermore, we introduce a coordinate scaling factor $e^{-\frac{r^2}{2}}$ to control the severity of its distortion with Eq. (15),

$$(x'', y'') = \left(x' e^{-\frac{r^2}{2}}, y' e^{-\frac{r^2}{2}} \right), \quad (15)$$

where r is the radial distance of the pixel (x', y') to the image center [53].

This mapping function introduces fisheye distortion, which manifests itself as an effect of “squeezing” the content towards the perimeter of the fisheye image. Different from the distortion in real-world fisheye images, which involves various camera parameters or lens distortion parameters, the mapping function we adopted is not camera or lens specific.

Synthesized Datasets. In the source datasets, the regions of faces usually appear around the center of image. However, in real applications, the region of face may appear elsewhere in a photo or video frame. The face region subjects neglectable geometric distortion if it is very close to the center of fisheye image. The degree of distortions increases as the face region moves towards the perimeter. To reflect the geometric distortion induced at different locations, a good fisheye dataset must have sufficient samples of which the face region may appear at all possible locations.

For each rectilinear image in the source datasets, we represented the location of face by the center of its bounding box. Then we created a canvas and replaced a canvas region with a facial region from the rectilinear image. In our work, the width and height of the canvas were both set to 5 times of the face bounding box. Next, we set up a polar coordinate system as the fisheye distortion is usually circularly symmetrical with respect to the image center [62]. The polar point was set at the canvas center and the angular angle increases counterclockwise. The radial distance of face was normalized by half the width of canvas and was sampled from the uniform distribution (0, 0.8). The polar angle of face was sampled from the uniform distribution (-180, 180). Finally, we transformed the canvas using the mapping function in Eqs. (14) and (15) to create the effect of fisheye distortion. The corresponding bounding box and landmarks were also converted into the fisheye image coordinate system. Once the

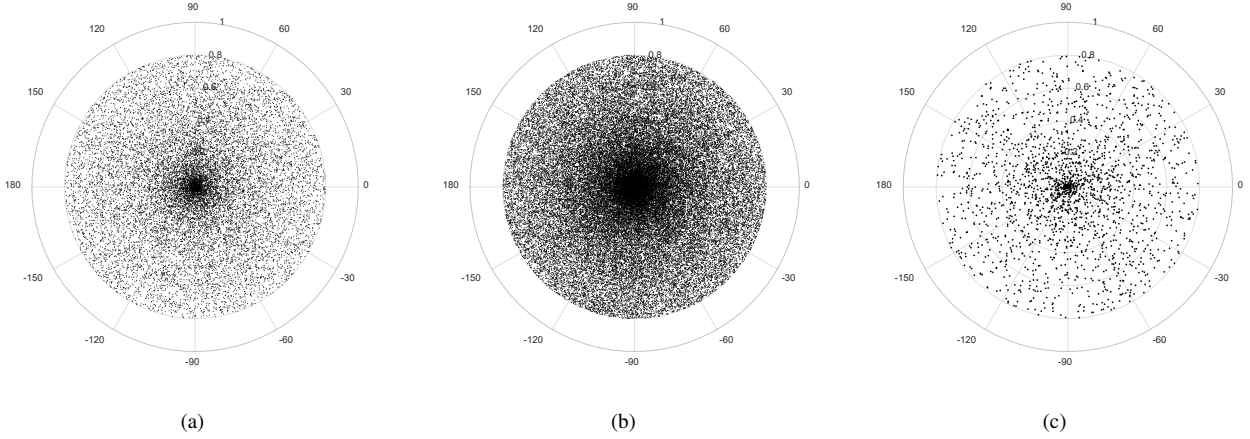


Fig. 6. The distribution of the face locations in polar coordinates for different fisheye datasets. (a) BIWI-360, (b) 300WP-LP-360, and (c) AFLW2000-360.

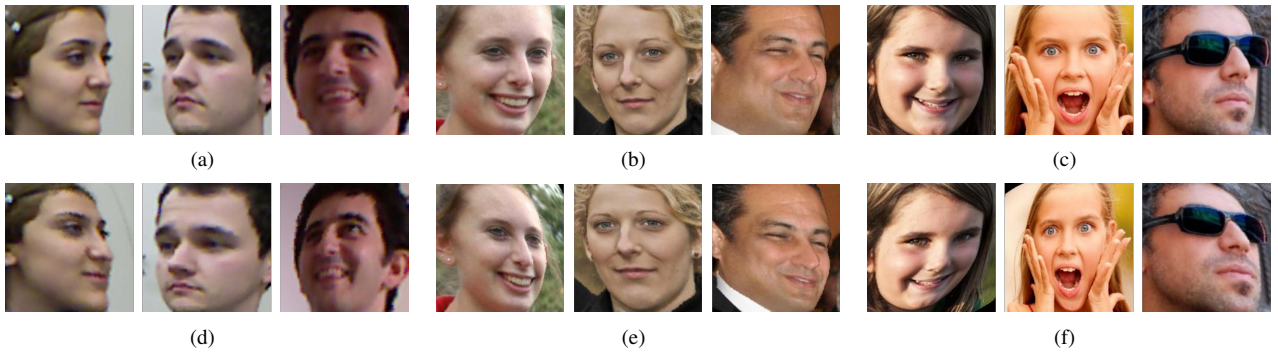


Fig. 7. Examples of face images from different datasets. The upper row represents original (rectilinear) face images from different datasets (a) BIWI, (b) 300W-LP, (c) AFLW2000, and the bottom row represents the corresponding fisheye face images from the synthesized datasets (d) BIWI-360, (e) 300W-LP-360, (f) AFLW2000-360.

coordinate transformation was completed, we cropped the face areas from the resulting fisheye image.

We performed the above operations on the BIWI, 300W-LP, and AFLW2000 datasets, and correspondingly obtained three synthesized fisheye datasets with varying face locations in the images. The distributions of the face locations in polar coordinates for these synthetic fisheye datasets are shown in Fig. 6. Each dot in these maps represents the face location in polar coordinates of each sample in the corresponding datasets. In order to have an intuitive impression on the synthesized fisheye datasets, we show some samples of face image from the original and the corresponding synthetic datasets in Fig. 7.

B. Implementations and Evaluation Metrics

In all our experiments, the proposed method was implemented on the Pytorch platform. Cuda 10.2 was also employed to speed up model training. Our network was trained for 25 epochs using Adam optimization [63] with an initial learning rate of 10^{-4} , $\beta_1 = 0.9$, and $\beta_2 = 0.9$. The learning rate was reduced at the 10^{th} epoch and the 20^{th} epoch by a factor of 0.5 each. Each color channel was normalized using the mean and standard deviation of ImageNet before training and testing. We set the value of α_{task} for the MSE loss of estimated normalized radial distance to 100, i.e., $\alpha_\rho = 100$, while the values of α_{task} for the MSE losses of estimated pitch, yaw,

roll, and polar angle to 1. Face detection was performed using RetinaFace [64].

Fig. 8 shows the overall loss and MAE losses of training and validation obtained in three experiments on BIWI-360 datasets. In these experiments, we randomly split the BIWI-360 dataset into 70% (16 videos) for training and 30% (8 videos) for validation. Fig. 8 indicated the proposed network converged after around 20 epochs.

Two evaluation metrics were used in all experiments. One was the average error in the yaw, pitch, and roll angles in degrees. The other metric was the mean absolute error (MAE), which is expressed in Eq. (16).

$$\text{MAE} = \frac{1}{3N} \sum_{k=1}^N \left(|\alpha_k - \hat{\alpha}_k| + |\beta_k - \hat{\beta}_k| + |\gamma_k - \hat{\gamma}_k| \right), \quad (16)$$

where N is the number of samples in the testing dataset, $(\alpha_k, \beta_k, \gamma_k)$ are the ground truth of the Euler angles and $(\hat{\alpha}_k, \hat{\beta}_k, \hat{\gamma}_k)$ are the estimated Euler angles.

C. Comparison with two-stage approaches

In this subsection, we compared the performance of our proposed method with the two-stage approaches. The two-stage approaches for comparison employed the progressively complementary network (PCN) [65] to rectify the fisheye

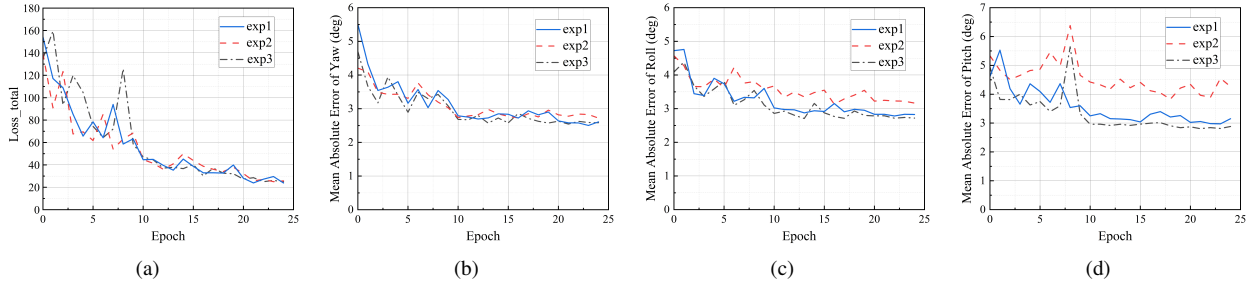


Fig. 8. The overall loss and MAE losses of the proposed network in training and validation on the BIWI-360 dataset. Experiments were repeated three times. (a) The overall loss of the proposed network in training; (b) the MAE loss of estimated yaw in validation; (c) the MAE loss of estimated roll in validation; (d) the MAE loss of estimated pitch in validation.

TABLE II

AVERAGE ERROR COMPARISONS ON THE AFLW2000-360 DATASET. OUR NETWORK WAS TRAINED ON THE 300W-LP-360 DATASET WHILE OTHERS WERE TRAINED ON THE 300W-LP DATASET.

Method	Yaw	Pitch	Roll	MAE
Hopenet ($\alpha = 1$) [33] + PCN	6.31°	8.20°	6.20°	6.90°
Hopenet ($\alpha = 2$) [33] + PCN	6.68°	8.36°	6.10°	7.05°
FSA-Net [34] + PCN	4.97°	7.66°	5.04°	5.89°
Li et al. [35] + PCN	5.39°	7.68°	5.23°	6.10°
6DRepNet [66] + PCN	5.57°	8.12°	5.95°	6.55°
Ours	3.65°	5.85°	4.17°	4.55°

TABLE III

AVERAGE ERROR COMPARISONS ON THE BIWI-360 DATASET. OUR NETWORK WAS TRAINED ON THE 300W-LP-360 DATASET WHILE OTHERS WERE TRAINED ON THE 300W-LP DATASET.

Method	Yaw	Pitch	Roll	MAE
Hopenet ($\alpha = 1$) [33] + PCN	6.21°	9.84°	4.26°	6.77°
Hopenet ($\alpha = 2$) [33] + PCN	6.98°	7.82°	4.32°	6.37°
FSA-Net [34] + PCN	6.37°	8.34°	4.28°	6.33°
Li et al. [35] + PCN	6.77°	8.13°	4.50°	6.47°
6DRepNet [66] + PCN	5.76°	8.46°	4.74°	6.32°
Ours	4.53°	7.24°	3.73°	5.16°

image first, and then estimated the head pose from the rectified image with the state-of-the-art HPE methods for rectilinear images. PCN was adopted as it is independent of the camera parameters and can generate an undistorted image from the fisheye image directly [65]. However, if the rectification error is very large, it may be the reason for the head pose error. We used the well-trained models of the two-stage approaches as reported in their corresponding publications. The estimation results obtained by using the two-stage approaches are reported as “original methods + PCN” in our experiments.

Our network was trained on the 300W-LP-360 dataset while others were trained on the 300W-LP dataset. For training our network, the batch size was set to 128, and the weights λ_1 and λ_2 in Eq. (13) were set to 5 and 0.001, respectively. In testing, our network used the fisheye image as input, while others used the rectified fisheye image as input. Testing was performed on two datasets: AFLW2000-360 and BIWI-360 datasets. The experimental results of testing on the AFLW2000-360 dataset are shown in Table II. The experimental results of testing on the BIWI-360 dataset are shown in Table III. Fig. 9 shows the estimated head pose on some samples from the BIWI-360 and AFLW2000-360 datasets.

As shown in Tables II and III, our proposed network obtained much lower average errors in yaw, pitch, roll angles, and MAE than the two-stage methods. Experiments demonstrated that the proposed method achieved better performance compared to other two-stage approaches.

Further investigation found our proposed method obtained large estimation error for some samples. We present some of these samples in Fig. 10. All these hard samples involve very

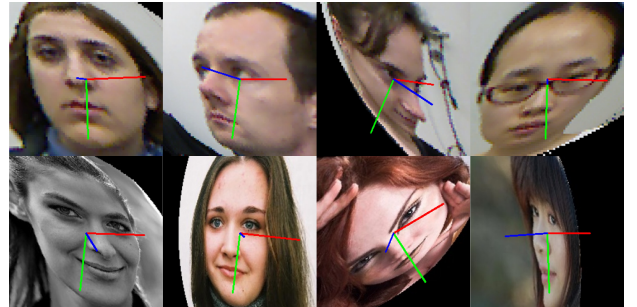


Fig. 9. Estimated head poses for examples from the BIWI-360 (the top line) and the AFLW2000-360 (the bottom line).

large head pose. They indicate the limited performance of our proposed method in dealing with images with large head pose. We argue that two reasons result in the large MAE for samples with large head pose. One reason is large head pose leads to the missing feature of head pose, which decreases the efficiency of the representation of head pose feature. The other reason is lacking of samples with large head pose in the training set, which limits the knowledge about large head pose the proposed model learned from the training set.

D. Comparison with one-stage approaches

Head pose estimators designed for rectilinear image can be retrained to deal with fisheye distortion as the one-stage approaches. We retrained some existing state-of-the-art methods for comparison, and cited them as “original methods + retrain” in experiment. We used the same configurations of the compared methods, including model architectures, loss func-



Fig. 10. Examples with large MAE in estimation. Our network was trained on the 300W-LP-360, and the test samples were from the AFLW2000-360 (the top line) and the BIWI-360 (the bottom line).

tion, and hyper-parameters, as reported in their corresponding publications. The experimental results are listed in Tables IV to VI.

Tables IV to VI show the average errors in degrees with different experimental protocols. We followed the experimental protocol proposed in [67] and randomly split the BIWI-360 dataset into 70% (16 videos) for training and 30% (8 videos) for testing to obtain the results shown in Table IV. We repeated this process three times and reported the average measurement error. In this protocol, the batch size was set to 32, and the weights λ_1 and λ_2 in Eq. (13) were set to 10 and 0.001, respectively. We also performed training on the 300W-LP-360 dataset while testing on the AFLW2000-360 dataset and the BIWI-360 dataset, respectively. We followed the same settings as Hopenet [33] and only considered samples whose Euler angles were within the range of $[-99^\circ, 99^\circ]$. In this protocol, the batch size was set to 128, and the weights λ_1 and λ_2 in Eq. (13) were set to 5 and 0.001, respectively. Results of these experiments are shown in Tables V and VI.

TABLE IV
COMPARISON OF AVERAGE ERRORS USING THE BIWI-360 DATASET FOR TRAINING AND TESTING.

Method	Yaw	Pitch	Roll	MAE
Hopenet ($\alpha = 1$) [33] + retrain	2.99°	3.50°	3.26°	3.25°
FSA-Net [34] + retrain	4.08°	5.14°	4.26°	4.50°
Li et al. [35] + retrain	3.39°	4.60°	3.35°	3.78°
6DRepNet [66] + retrain	3.17°	3.56°	3.25°	3.32°
Ours	2.66°	3.11°	2.87°	2.88°

Table IV shows our method obtained the best performance among all compared methods when used the BIWI-360 dataset for training and testing. The training for Tables V and VI were performed on the 300W-LP-360 dataset. Although our method had a slightly higher error than the retrained 6DRepNet in the testing on the AFLW2000-360 dataset, our method requires a much smaller model size than 6DRepNet. Moreover, when tested on the BIWI-360 dataset (Table VI), our method performed better in estimating yaw, roll angles, and MAE, despite having a comparable error in pitch angle compared to the retrained 6DRepNet.

The above results show that our approach achieved better or comparable performance compared to other one-stage ap-

TABLE V
COMPARISON OF AVERAGE ERRORS USING THE 300W-LP-360 DATASET FOR TRAINING AND THE AFLW2000-360 DATASET FOR TESTING.

Method	Yaw	Pitch	Roll	MAE
Hopenet ($\alpha = 1$) [33] + retrain	3.60°	6.05°	4.42°	4.69°
Hopenet ($\alpha = 2$) [33] + retrain	3.59°	5.96°	4.49°	4.68°
FSA-Net [34] + retrain	5.61°	7.16°	5.70°	6.16°
Li et al. [35] + retrain	4.19°	6.29°	4.49°	4.99°
6DRepNet [66] + retrain	3.47°	5.48°	3.69°	4.21°
Ours	3.65°	5.85°	4.17°	4.55°

TABLE VI
COMPARISON OF AVERAGE ERRORS USING THE 300W-LP-360 DATASET FOR TRAINING AND THE BIWI-360 DATASET FOR TESTING.

Method	Yaw	Pitch	Roll	MAE
Hopenet ($\alpha = 1$) [33] + retrain	5.04°	7.22°	3.77°	5.35°
Hopenet ($\alpha = 2$) [33] + retrain	5.22°	7.51°	3.77°	5.50°
FSA-Net [34] + retrain	6.18°	9.53°	4.58°	6.76°
Li et al. [35] + retrain	5.07°	7.99°	4.10°	5.72°
6DRepNet [66] + retrain	5.26°	6.95°	3.93°	5.38°
Ours	4.53°	7.24°	3.73°	5.16°

proaches. As we used head location estimation as the auxiliary task for head pose estimation, these experiments demonstrate the knowledge of head location improves the performance of our network for head pose estimation.

E. Processing Speed

We also conducted an experiment to evaluate the processing speed of our network. We compared our network with two-stage methods and one-stage methods on a GPU platform. The GPU platform employed a single NVIDIA GeForce TITAN X. The version of CUDA library was 10.2. For a fair comparison, all methods were tested on the same target platform. Each method was evaluated 100 times. The measured average runtime is shown in Tables VII and VIII.

Tables VII and VIII show our proposed method estimates head pose from a facial image in nearly 14.1ms (or 71fps) with the 28.92M parameters. As shown in Table VII, our method runs faster and uses a smaller number of parameters compared with the two-stage methods because our network does not require a separate image rectification process. Table VIII shows that our network runs slower than other one-stage approaches because our network includes a location branch and additional computational cost. Nevertheless, the processing speed of our method is good enough for real-world applications in most cases.

F. Ablation Study

Experiments were conducted to show the effectiveness of our proposed head location guided head pose estimation on fisheye distortion. All experiments used the BIWI-360 dataset for training (70%) and testing (30%). We repeated this process three times and reported the average measurement error. The

TABLE VII

THE COMPARISON OF PROCESSING SPEED AND NUMBER OF PARAMETERS FOR OUR NETWORK AND TWO-STAGE METHODS.

Method	Inference time (ms)	Speed (fps)	#Params
Hopenet [33] + PCN	55.6	18	59.56M
FSA-Net [34] + PCN	52.6	19	36.81M
Li et al. [35] + PCN	47.6	21	35.85M
6DRepNet [66] + PCN	48.5	21	74.96M
Ours without LFEM ¹	12.5	80	24.20M
Ours	14.1	71	28.92M

¹ LFEM stands for location feature extraction module.

TABLE VIII

THE COMPARISON OF PROCESSING SPEED AND NUMBER OF PARAMETERS FOR OUR NETWORK AND ONE-STAGE METHODS.

Method	Inference time (ms)	Speed (fps)	#Params
Hopenet [33] + retrain	15.2	66	23.92M
FSA-Net [34] + retrain	11.9	84	1.71M
retrain	8.1	123	0.21M
6DRepNet [66] + retrain	18.9	53	43.75M
Ours without LFEM ¹	12.5	80	24.20M
Ours	14.1	71	28.92M

¹ LFEM stands for location feature extraction module.

experimental results are shown in Table IX. In Table IX, \mathcal{L}_ρ denotes the loss function for normalized radial distance used in training. \mathcal{L}_θ represents the loss function for polar angle used in training.

To further verify the impact of location guidance, we performed an ablation study in every possible configuration, i.e., the network is equipped with or without the location feature extraction module, the location supervisions for polar angle \mathcal{L}_θ , and the location supervisions for normalized radial distance \mathcal{L}_ρ . In the case no location feature extraction module was used, we connected the output of backbone module to the input of location estimation module directly and cut off the connection between the backbone module and the adder. That means the location estimation module estimated the location of head position directly from F_{basic} . The loss of the estimation of head center location was backpropagated for the parameter updating of backbone module and location estimation module. If the loss for head center location was not employed in ablation study, it was not backpropagated for parameter updating. The experimental results are listed in Table IX.

The first four rows of Table IX show that using the guidance of head location, although without the location feature extraction module, helped improve accuracy of the network. The first and fifth rows of Table IX show that the location feature extraction module without location supervision would not be beneficial in improving the estimation accuracy of the network. The last three rows of Table IX show that the combination

TABLE IX

THE EFFECTIVENESS OF THE LOCATION GUIDANCE OF HEAD FOR HPE. THE BIWI-360 DATASET WAS USED FOR TRAINING (70%) AND TESTING (30%). “✓” AND “-” INDICATE WITH OR WITHOUT THE CORRESPONDING MODULE/SUPERVISION.

Location feature extraction module	Location supervision		Yaw	Pitch	Roll	MAE
	\mathcal{L}_ρ	\mathcal{L}_θ				
-	-	-	2.72°	3.41°	3.06°	3.06°
-	-	✓	2.71°	3.25°	2.98°	2.98°
-	✓	-	2.59°	3.36°	2.94°	2.96°
-	✓	✓	2.66°	3.31°	3.03°	3.00°
✓	-	-	2.74°	3.28°	3.14°	3.06°
✓	-	✓	2.55°	3.27°	2.97°	2.93°
✓	✓	-	2.56°	3.25°	2.95°	2.92°
✓	✓	✓	2.66°	3.11°	2.87°	2.88°

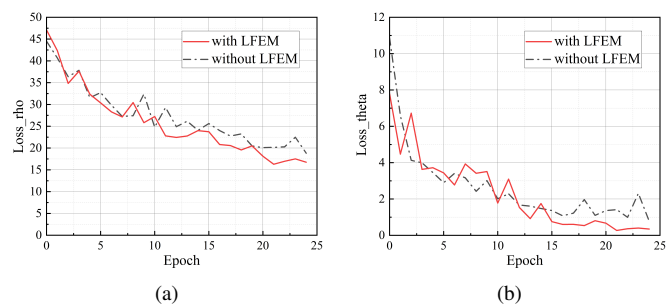


Fig. 11. The training losses (defined in Eq. 12) of the estimated (a) normalized radial distance ρ and (b) polar angle θ with and without the location feature extraction module. LFEM stands for location feature extraction module.

of head location supervision and location feature extraction module contributed to more accurate estimation results. The first and last rows of Table IX show that the network using both location feature extraction module and head location had an across-the-board improvement and 5.88% lower MAE compared with the network without the knowledge of head location.

We performed more ablation study to evaluate the help of the location feature extraction module in the estimation of head location. As shown in Fig. 11, the network with the location feature extraction module obtained faster decreased loss values in training for the estimation of the normalized radial distance ρ and polar angle θ , compared with the network without the location feature extraction module.

We also conducted an ablation study using different backbone networks, including VGG11 and MobileNet V3, to evaluate the performance of the proposed framework. Table X shows the average errors obtained by our proposed network using different backbones. Configurations of this ablation study were the same as described in Subsection IV. D. Results in Table X show our proposed network obtained the best average errors when using ResNet50 as the backbone in every experiment configuration.

G. Studies on Hyperparameters

Experiments were conducted to explore the sensitivity of our proposed approach for the hyperparameters λ_1 and λ_2 .

TABLE X
PERFORMANCE OF THE PROPOSED NETWORK USING DIFFERENT
BACKBONES

Training set	Test set	Backbone	Yaw	Pitch	Roll	MAE
BIWI-360	BIWI-360	ResNet50	2.66°	3.11°	2.87°	2.88°
		VGG11	2.92°	3.39°	3.23°	3.18°
		MobileNet v3	2.98°	3.61°	3.39°	3.32°
300W-LP -360	AFLW2000 -360	ResNet50	3.65°	5.85°	4.17°	4.55°
		VGG11	3.58°	6.26°	4.35°	4.73°
		MobileNet v3	4.03°	6.29°	4.50°	4.94°
300W-LP -360	BIWI-360	ResNet50	4.53°	7.24°	3.73°	5.16°
		VGG11	5.76°	8.30°	4.22°	6.10°
		MobileNet v3	5.54°	7.11°	4.13°	5.59°

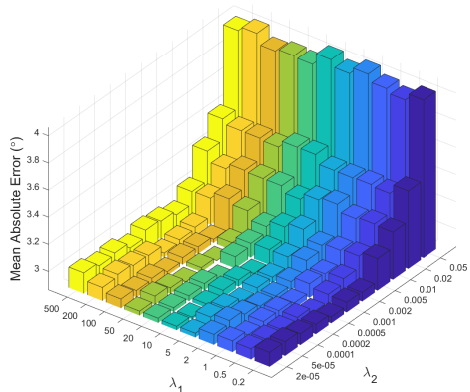


Fig. 12. The mean absolute error of Euler angles (in degrees) with different λ_1 and λ_2 on the BIWI-360 dataset. We use logarithmic scales to show the values of λ_1 and λ_2 .

We tested the performance of our proposed network by setting each of λ_1 and λ_2 to 11 different values. Specifically, we set λ_1 to 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100, 200, 500, and λ_2 to 0.00002, 0.00005, 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05 in the experiment. All experiments used the BIWI-360 dataset for training (70%) and testing (30%). We repeated this process three times and reported the average measurement error. The experimental results are visualized in Fig. 12.

The experimental results show that the best performance was achieved by setting λ_1 between 10 and 20, and λ_2 between 0.0005 and 0.001. When λ_1 and λ_2 were set to 10 and 0.001, the obtained mean absolute error of Euler angles was 2.88°.

H. Real-World Head Pose Estimation

To evaluate the performance of our proposed method on real fisheye images, we constructed a real-world dataset with fisheye images captured by a calibrated fisheye camera (JR®HF900). The focal length and FOV of this camera are 1.1mm and 210°, respectively. The resolution of the captured fisheye image is 1080×1080p. The dataset contains 70 sequences of 10 different people (7 men and 3 women, 5 wearing glasses) recorded while sitting or standing about 0.5



Fig. 13. Some samples of real-world fisheye images captured by JR®HF900.

meters away from the fisheye camera. Some sample frames of sequences captured by the fisheye camera are shown in Fig. 13. The subjects rotated their heads to various positions. We captured videos of each subject from seven different angles, so that the subjects could appear in all possible locations within the fisheye images. More specifically, we kept the camera’s optical axis to be approximately horizontal, and then used a camera gimbal to turn the fisheye camera on the horizontal plane so that the angle of the optical axis was at 0°, ±15°, ±30°, and ±45° to the direction of subject’s face. For each fisheye image, we employed RetinaFace [64] to detect the face region and discarded the fisheye images when face detection failed to detect the face.

To annotate the head pose in the real-world fisheye image, we used the Camera Calibrator App [68] provided by MATLAB R2021b to estimate the parameters of our fisheye camera, and then used these camera parameters to rectify the image captured by the fisheye camera. Although this process was tedious and time-consuming, the obtained camera parameters helped us decrease the image distortion induced from lens of the fisheye camera and produce precise rectified image from the fisheye. Finally, an excellent rectilinear-image-based HPE method [24] was employed to estimate the head pose from the rectified image. The Euler angles estimated by the HPE method was regarded as the ground truth of the head pose in the real-world fisheye image. As a result, the real-world dataset contains 21,815 fisheye images, annotated with head pose in Euler angles. The head pose range covers about ±90° for yaw, ±60° for pitch, and ±45° for roll.

We evaluated our proposed method on the constructed real-world dataset and provides qualitatively demonstrated the performance of our method in Fig. 14. The upper row of Fig. 14 shows the visualized ground truth of some real-world samples and the bottom row shows the head poses estimated by our method. These samples were captured from different angles and include all subjects. This visualized comparison

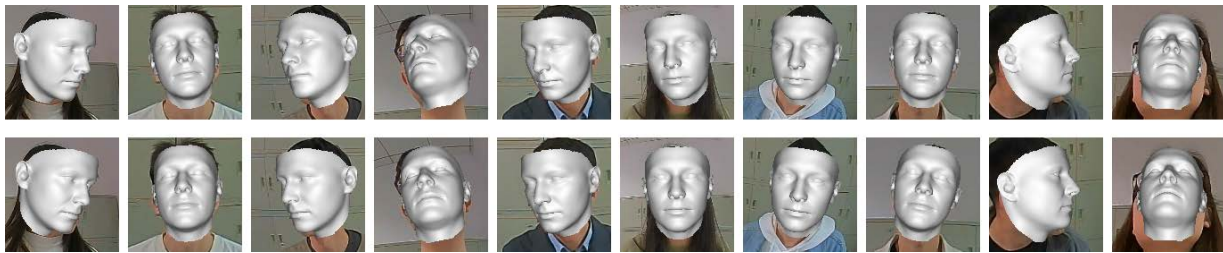


Fig. 14. Comparison of the visualized ground truth of real-world samples (the upper row) and the head pose estimated by our proposed method (the bottom row). Head poses are visualized by the tools from [69].

TABLE XI
COMPARISON OF OUR PROPOSED METHOD AND TWO-STAGE METHODS ON REAL-WORLD FISHEYE IMAGES IN THE CASE NO CAMERA PARAMETERS ARE PROVIDED.

Method	Yaw	Pitch	Roll	MAE
Hopenet ($\alpha = 1$) [33] + PCN	31.04°	17.57°	9.52°	19.37°
Hopenet ($\alpha = 2$) [33] + PCN	28.97°	17.71°	10.74°	19.14°
FSA-Net [34] + PCN	15.85°	18.66°	8.11°	14.21°
Li et al. [35] + PCN	26.64°	17.49°	8.33°	17.49°
Ours	9.23°	8.12°	6.37°	7.91°

shows that the poses estimated by our method are close to the ground truth, indicating the effectiveness of our approach with fisheye images of faces appearing at different locations.

We also compared the performances of our proposed method with two-stage methods on real-world fisheye images in the case that no camera parameters are provided. These experimental results are listed in Table XI. Experimental results show that our proposed method achieved better performance compared to the two-stage approaches in the case that no camera parameters are provided.

It is worth emphasizing that all experiments we performed, including on the synthetic and real-world datasets, assume a horizontal optical axis for the camera, which means the fisheye images in these datasets are mostly collected from a frontal view. In another common application scenario, ceiling-mounted fisheye cameras are used to capture top-view fisheye images [20], [22].

Although our proposed algorithm does not require the input image to be a front-view fisheye image, it will meet great challenges when the input is a top-view fisheye image for two reasons. First, the proposed method can hardly detect the presence of human face from the top view of a person. We employ a process of face detection for frontal view images to help the algorithm focus on the region of human face. For a fisheye image taken by a ceiling-mounted camera, the performance of the employed face detection algorithm will be affected dramatically. It may make the proposed method miss the head region appeared in the fisheye image. Second, the proposed method has little knowledge of the mapping between head pose and the top-view image of human.

As we use the synthesized fisheye datasets for the network training, while the synthesized fisheye datasets come from the datasets acquired by the camera with horizontal or nearly

horizontal optical axis, the trained network only learned the knowledge about the head pose based on the fisheye image from the frontal view. We believe the knowledge of the proposed method comes from the datasets used in training. Datasets composed of top-view fisheye images together with the corresponding labels of head pose would help our proposed algorithm learn the knowledge of head pose from fisheye images taken by ceiling-mounted cameras.

The proposed method focuses on estimating head pose directly from fisheye image without the process of rectification or calibration. Although the proposed method obtained promising performance on synthetic datasets, it does not specifically discuss the issue of face detection from fisheye image in real-world applications. Technique of face detection helps HPE algorithm exclude other parts of the body or the scene from the analysis. Unsuccessful method of face detection usually leads to a failure of head pose estimation. Many excellent algorithms achieved promising results in recent years as they well addressed the multi-scale problems appeared in the domain of face detection [70] [71]. Although these methods are developed for rectilinear image, they of course enlighten the research on face detection for fisheye image and contribute the improvement of head pose estimation.

V. CONCLUSION

The rapid growth of the fisheye camera market has made the cognition and comprehension techniques for fisheye image an emerging research domain in recent years. Estimating head pose in fisheye image helps artificial systems understand the intention of humans directly from an ultra-wide field-of-view. To the best of our knowledge, there is no reported research of HPE specifically designed for fisheye image.

Head pose estimation from the fisheye image remains a challenging issue because of serious fisheye distortion in the peripheral region of the image. We propose an approach using the knowledge of head location to reduce the impact of radial distortion for head pose estimation. The proposed network does not require the process of rectification or calibration. It estimates the head pose by a multi-task learning of head location and head pose. Experimental results show that the knowledge of the head location in the image contributes to the improved HPE accuracy. Our network outperforms the two-stage methods on two synthetic fisheye datasets and the one-stage methods on three synthetic fisheye datasets. Additionally, our network has a faster speed compared with the two-stage methods that involve a separate image rectification process.

This work is believed to fill the gap between the potential requirement of HPE for fisheye image and the previous techniques.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (62173353, 62171207), Science and Technology Program of Guangzhou, China (202007030011, 2023B03J1327), and Guangdong Science and Technology Program (2023A1111120012).

REFERENCES

- [1] K. Khan, R. U. Khan, R. Leonardi, P. Migliorati, and S. Benini, "Head pose estimation: A survey of the last ten years," *Signal Processing: Image Communication*, vol. 99, p. 116479, 2021.
- [2] C.-W. Chen and H. Aghajan, "Multiview social behavior analysis in work environments," in *2011 Fifth ACM/IEEE International Conference on Distributed Smart Cameras*. IEEE, 2011, pp. 1–6.
- [3] N. Alioua, A. Amine, A. Rogozan, A. Bensrhair, and M. Rziza, "Driver head pose estimation using efficient descriptor fusion," *EURASIP Journal on Image and Video Processing*, vol. 2016, no. 1, pp. 1–14, 2016.
- [4] C. Bisogni, L. Cascone, M. Nappi, and C. Pero, "Iot-enabled biometric security: Enhancing smart car safety with depth-based head pose estimation," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 155, pp. 1–24, 2024.
- [5] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10034–10044.
- [6] B. Benfold and I. D. Reid, "Guiding visual surveillance by tracking human attention," in *British Machine Vision Conference*, vol. 2, no. 6, 2009, p. 7.
- [7] C. George, D. Buschek, A. Ngao, and M. Khamis, "Gazeroomlock: Using gaze and head-pose to improve the usability and observation resistance of 3d passwords in virtual reality," in *Augmented Reality, Virtual Reality, and Computer Graphics*, L. T. De Paolis and P. Bourdot, Eds. Cham: Springer International Publishing, 2020, pp. 61–81.
- [8] K. Wang, R. Zhao, and Q. Ji, "Human computer interaction with head pose, eye gaze and body gestures," in *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, 2018, pp. 789–789.
- [9] J.-M. Odobez and S. Ba, "Cognitive and unsupervised map adaptation approach to the recognition of the focus of attention from head pose," in *2007 IEEE International Conference on Multimedia and Expo*, 2007, pp. 1379–1382.
- [10] A. F. Abate, C. Bisogni, A. Castiglione, and M. Nappi, "Head pose estimation: An extensive survey on recent techniques and applications," *Pattern Recognition*, vol. 127, p. 108591, 2022.
- [11] P. Hansen, P. Corke, W. Boles, and K. Daniilidis, "Scale-invariant features on the sphere," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.
- [12] J. Cruz-Mota, I. Bogdanova, B. Paquier, M. Bierlaire, and J.-P. Thiran, "Scale invariant feature transform on the sphere: Theory and applications," *International journal of computer vision*, vol. 98, pp. 217–241, 2012.
- [13] C. Demonceaux, P. Vasseur, and Y. Fougerolle, "Central catadioptric image processing with geodesic metric," *Image and Vision Computing*, vol. 29, no. 12, pp. 840–849, 2011.
- [14] K. K. Delibasis, "Efficient implementation of gaussian and laplacian kernels for feature extraction from ip fisheye cameras," *Journal of Imaging*, vol. 4, no. 6, p. 73, 2018.
- [15] S. V. Georgakopoulos, K. Kottari, K. Delibasis, V. P. Plagianakos, and I. Maglogiannis, "Pose recognition using convolutional neural networks on omni-directional images," *Neurocomputing*, vol. 280, pp. 23–31, 2018.
- [16] W. Pang, Q. Gao, Y. Zhao, Z. Ju, and J. Hu, "Basicnet: Lightweight 3-d hand pose estimation network based on biomechanical structure information for dexterous manipulator teleoperation," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 16, no. 2, pp. 448–457, 2024.
- [17] G. Wang, H. Zeng, Z. Wang, Z. Liu, and H. Wang, "Motion projection consistency-based 3-d human pose estimation with virtual bones from monocular videos," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 2, pp. 784–793, 2023.
- [18] Q. Wu, Y. Zhang, L. Zhang, and H. Yu, "Parallel self-attention and spatial-attention fusion for human pose estimation and running movement recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 16, no. 1, pp. 358–368, 2024.
- [19] Z. Zhu, D. Zhang, C. Chi, M. Li, and D.-J. Lee, "A complementary dual-branch network for appearance-based gaze estimation from low-resolution facial image," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 3, pp. 1323–1334, 2023.
- [20] D. Tseng, T. I. of Computer Science, C. U. Information Engineering, C. Chen, and C. Tseng, "Automatic detection and tracking in multi-fisheye cameras surveillance," *International Journal of Computer and Electrical Engineering*, vol. 9, no. 1, pp. 370–383, 2017.
- [21] Y. Ye, K. Yang, K. Xiang, J. Wang, and K. Wang, "Universal semantic segmentation for fisheye urban driving images," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 648–655.
- [22] K. N. Kottari, K. K. Delibasis, and I. G. Maglogiannis, "Real-time fall detection using uncalibrated fisheye cameras," *IEEE transactions on cognitive and developmental systems*, vol. 12, no. 3, pp. 588–600, 2019.
- [23] D. Zhu, Y. Chen, D. Zhao, Y. Zhu, Q. Zhou, G. Zhai, and X. Yang, "Multiscale brain-like neural network for saliency prediction on omni-directional images," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 507–518, 2021.
- [24] C.-Y. Yang and H. H. Chen, "Efficient face detection in the fisheye image domain," *IEEE Transactions on Image Processing*, vol. 30, pp. 5641–5651, 2021.
- [25] Y.-C. Lo, C.-C. Huang, Y.-F. Tsai, I.-C. Lo, A.-Y. A. Wu, and H. H. Chen, "Face recognition for fisheye images," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 146–150.
- [26] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1513–1520.
- [27] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International journal of computer vision*, vol. 107, pp. 177–190, 2014.
- [28] 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 *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1021–1030.
- [29] H. Zhang, Q. Li, and Z. Sun, "Joint voxel and coordinate regression for accurate 3d facial landmark localization," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2202–2208.
- [30] M. Ahmad, D. Ai, G. Xie, S. F. Qadri, H. Song, Y. Huang, Y. Wang, and J. Yang, "Deep belief network modeling for automatic liver segmentation," *IEEE Access*, vol. 7, pp. 20 585–20 595, 2019.
- [31] I. Hirra, M. Ahmad, A. Hussain, M. U. Ashraf, I. A. Saeed, S. F. Qadri, A. M. Alghamdi, and A. S. Alfakeeh, "Breast cancer classification from histopathological images using patch-based deep learning modeling," *IEEE Access*, vol. 9, pp. 24 273–24 287, 2021.
- [32] S. F. Qadri, L. Shen, M. Ahmad, S. Qadri, S. S. Zareen, and S. Khan, "Op-convnet: A patch classification-based framework for ct vertebrae segmentation," *IEEE Access*, vol. 9, pp. 158 227–158 240, 2021.
- [33] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 2074–2083.
- [34] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang, "Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1087–1096.
- [35] X. Li, D. Zhang, M. Li, and D.-J. Lee, "Accurate head pose estimation using image rectification and a lightweight convolutional neural network," *IEEE Transactions on Multimedia*, 2022.
- [36] H.-W. Hsu, T.-Y. Wu, S. Wan, W. H. Wong, and C.-Y. Lee, "Quatnet: Quaternion-based head pose estimation with multiregression loss," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1035–1046, 2019.
- [37] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner, "img2pose: Face alignment and detection via 6dof, face pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7617–7627.

- [38] C. Geyer and K. Daniilidis, "A unifying theory for central panoramic systems and practical implications," in *Computer Vision-ECCV 2000: 6th European Conference on Computer Vision Dublin, Ireland, June 26–July 1, 2000 Proceedings, Part II 6*. Springer, 2000, pp. 445–461.
- [39] X. Ying and Z. Hu, "Can we consider central catadioptric cameras and fisheye cameras within a unified imaging model," in *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004. Proceedings, Part I 8*. Springer, 2004, pp. 442–455.
- [40] J. Fan, J. Zhang, S. J. Maybank, and D. Tao, "Wide-angle image rectification: A survey," *International Journal of Computer Vision*, vol. 130, no. 3, pp. 747–776, 2022.
- [41] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.
- [42] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, 2017, pp. 17–24.
- [43] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019.
- [44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 770–778.
- [45] I. Rieger, T. Hauenstein, S. Hettenkofer, and J.-U. Garbas, "Towards real-time head pose estimation: Exploring parameter-reduced residual networks on in-the-wild datasets," in *Advances and Trends in Artificial Intelligence. From Theory to Practice: 32nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2019, Graz, Austria, July 9–11, 2019, Proceedings 32*. Springer, 2019, pp. 123–134.
- [46] M. Shao, Z. Sun, M. Ozay, and T. Okatani, "Improving head pose estimation with a combined loss and bounding box margin adjustment," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–5.
- [47] B. Huang, R. Chen, W. Xu, and Q. Zhou, "Improving head pose estimation using two-stage ensembles with top-k regression," *Image and Vision Computing*, vol. 93, p. 103827, 2020.
- [48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.
- [49] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [50] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3d face analysis," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 437–458, 2012.
- [51] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3d total solution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 78–92, 2019.
- [52] J. Fu, S. R. Alvar, I. Bajić, and R. Vaughan, "Fddb-360: Face detection in 360-degree fisheye images," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2019, pp. 15–19.
- [53] J. Fu, I. V. Bajić, and R. G. Vaughan, "Datasets for face and object detection in fisheye images," *Data in brief*, vol. 27, p. 104752, 2019.
- [54] Y.-H. Li, I.-C. Lo, and H. H. Chen, "Deep face rectification for 360° dual-fisheye cameras," *IEEE Transactions on Image Processing*, vol. 30, pp. 264–276, 2021.
- [55] T. Li, G. Tong, H. Tang, B. Li, and B. Chen, "Fisheyedet: A self-study and contour-based object detector in fisheye images," *IEEE Access*, vol. 8, pp. 71 739–71 751, 2020.
- [56] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.
- [57] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2879–2886.
- [58] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 386–391.
- [59] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 397–403.
- [60] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre *et al.*, "Xm2vtsdb: The extended m2vts database," in *Second international conference on audio and video-based biometric person authentication*, vol. 964. Citeseer, 1999, pp. 965–966.
- [61] M. Koestinger, 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 *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 2144–2151.
- [62] K. Zhao, C. Lin, K. Liao, S. Yang, and Y. Zhao, "Revisiting radial distortion rectification in polar-coordinates: A new and efficient learning perspective," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, p. 3552–3560, 2022.
- [63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015, pp. 1–13.
- [64] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5203–5212.
- [65] S. Yang, C. Lin, K. Liao, C. Zhang, and Y. Zhao, "Progressively complementary network for fisheye image rectification using appearance flow," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6348–6357.
- [66] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, "Toward robust and unconstrained full range of rotation head pose estimation," *IEEE Transactions on Image Processing*, vol. 33, pp. 2377–2387, 2024.
- [67] S. S. Mukherjee and N. M. Robertson, "Deep head pose: Gaze-direction estimation in multimodal video," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2094–2107, 2015.
- [68] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A toolbox for easily calibrating omnidirectional cameras," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2006, pp. 5695–5701.
- [69] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *European Conference on Computer Vision*. Springer, 2020, pp. 152–168.
- [70] S. D. Khan, H. Ullah, M. Ullah, N. Conci, F. A. Cheikh, and A. Beghdadi, "Person head detection based deep model for people counting in sports videos," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2019, pp. 1–8.
- [71] S. D. Khan and S. Basalamah, "Scale and density invariant head detection deep model for crowd counting in pedestrian crowds," *The Visual Computer*, vol. 37, pp. 2127–2137, 2021.



Bing Li received her B.Eng. degree from Sun Yat-sen University, China, in 2021. She is currently a postgraduate student in the school of Electronics and Information Technology, Sun Yat-sen University. Her research interests include head pose estimation and computer vision. E-mail: libing29@mail2.sysu.edu.cn.



Dong Zhang received his B.S.E.E. and M. S. degrees from Nanjing University, China, in 1999 and 2003, respectively, and Ph.D. degree from Sun Yat-sen University, China, in 2009. He is currently an associate professor in the school of Electronics and Information Technology, Sun Yat-sen University. His research interests include image processing, pattern recognition and information hiding.



Cheng Huang received his B.Eng. degree from Sun Yat-sen University, China, in 2021. He is currently a postgraduate student in the school of Electronics and Information Technology, Sun Yat-sen University. His research interests include group activity recognition and computer vision. E-mail: huangch63@mail2.sysu.edu.cn.



Yun Xian received the B.Eng. degree from the School of Information Science and Technology, Jinan University, Guangzhou, China, in 2021. She is currently pursuing the master's degree with the School of Information and Communication Engineering, Sun Yat-sen University. Her research interests include expression spotting and computer vision. E-mail: xiany7@mail2.sysu.edu.cn.



Ming Li received his Ph.D. in Electrical Engineering from University of Southern California in May 2013. He is currently an associate professor of the Data Science Research Center at Duke Kunshan University, a research scholar at the ECE department of Duke University, and the adjunct professor at Wuhan University. His research interests are in the areas of speech processing and multimodal behavior signal analysis with applications to human centered behavioral informatics notably in health, education and security.



Dah-Jye Lee received his M.S. and Ph.D. degrees in electrical engineering from Texas Tech University and MBA degree from Shenandoah University. He served in the machine vision industry for eleven years before joining Brigham Young University faculty in 2001. He is currently a professor and the director of the Robotic Vision Laboratory in the Electrical and Computer Engineering Department at BYU. His research focuses on artificial intelligence, high-performance visual computing, robotic vision, and visual inspection automation.