MFDNet: Collaborative Poses Perception and Matrix Fisher Distribution for Head Pose Estimation

Hai Liu, *Senior Member, IEEE*, Shuai Fang, Zhaoli Zhang, Duantengchuan Li, Ke Lin, and Jiazhang Wang

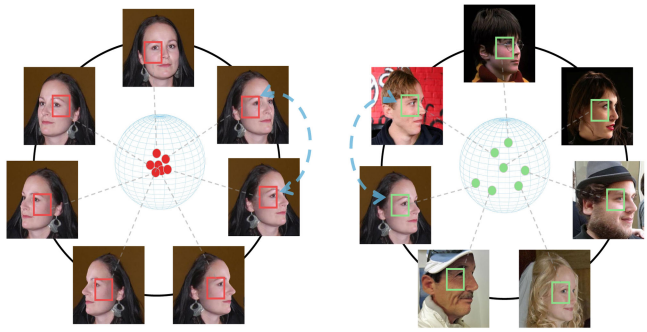
***Abstract*—Head pose estimation suffers from several problems, including low pose tolerance under different disturbances and ambiguity arising from common head pose representation. In this study, a robust three-branch model with triplet module and matrix Fisher distribution module is proposed to address these problems. Based on metric learning, the triplet module employs triplet architecture and triplet loss. It is implemented to maximize the distance between embeddings with different pose pairs and minimize the distance between embeddings with same pose pairs. It can learn a highly discriminate and robust embedding related to head pose. Moreover, the rotation matrix instead of Euler angle and unit quaternion is utilized to represent head pose. An exponential probability density model based on the rotation matrix (referred to as the matrix Fisher distribution) is developed to model head rotation uncertainty. The matrix Fisher distribution can further analyze the head pose, and its maximum likelihood obtained using singular value decomposition provides enhanced accuracy. Extensive experiments executed over AFLW2000 and BIWI datasets demonstrate that the proposed model achieves state-of-the-art performance in comparison with traditional methods.**

***Index Terms*—Head pose estimation, Triplet loss, Rotation matrix, Matrix Fisher distribution, Metric learning.**

I. Introduction

HEAD pose estimation (HPE) has become an important topic, and it is realized by detecting the direction of the face and judging the rotation of the head. HPE has been applied in various areas, such as emotion detection [1], gaze detection [2], human-computer interaction [3], and human behavior analysis [4], etc. Recently, HPE from RGB images has gradually become popular because it does not always require expensive devices, such as sensors, nor additional depth information. The majority of previous studies [5], [6] used the Perspective-n-Point method to determine the projection relationship of 2D facial landmarks with the 3D generic head model and obtain head poses. For the limitations in facial landmark detection, several competitive methods based on convolutional neural network (CNN) [7], [8] have been proposed, these methods directly predict head poses without facial landmarks. Despite these developments, two fundamental problems in HPE remain.

The first problem is that HPE is susceptible to disturbances, such as identity, illumination, and occlusion. The problem can be observed in Fig. 1, which we refer to as low pose tolerance. Taking several images with different and similar poses from Fig. 1 as examples, the facial features from the facial regions marked by the boxes are extracted using HopeNet [7]. Then a cluster analysis is conducted on these facial features. As shown in Fig. 1(a), the facial features marked by red boxes are highly similar, whereas in Fig. 1(b), these images with the same pose from different persons have different facial features. Specifically, the features closely related to the head pose may be unrelated to identity, and the similar facial features in the images with different poses may result in a confusing prediction for HPE. To address the low pose tolerance, the common features in images with the same pose should be explored. It is reasonable to believe that the multiple images with the same pose as an input can provide additional information on the head pose. Hence, the



(a) The same person with different poses (b) The same pose by different persons

Fig. 1. Head pose estimation suffers from low pose tolerance. (a) The images have the same identity but different poses. (b) The images have different identities and the same pose as one of the images in (a). The facial regions indicated by the boxes are fed into HopeNet for extracting the facial features. The images with different poses share more similar facial features than the images with the same pose.

Manuscript received January 2, 2021; revised April 17, 2021; accepted May 10, 2021. Date of publication May 19, 2021; date of current version May 11, 2022. This work was supported in part by the National Natural Science Foundation of China under Grants 62011530436, 62077020, 62005092, and 61875068, and the Fundamental Research Funds for the Central Universities under Grants CCNU20ZT017 and CCNU2020ZN008. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sebastian Knorr *(Corresponding author: Shuai Fang)*

Hai Liu, Zhaoli Zhang, and Duantengchuan Li are with the National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China (e-mail: hailiu0204@mail.ccnu.edu.cn; zl.zhang@mail.ccnu.edu.cn; dtclee1222@gmail.com).

Shuai Fang is with the National Engineering Laboratory For Educational Big Data, Central China Normal University, Wuhan 430079, China (e-mail: fibreggs@gmail.com).

Ke Lin is with the Control Science and Engineering, Harbin Institute of Technology, Shenzhen 150001, China (e-mail: sheldon.chris.lin@gmail.com).

Jiazhang Wang is with Northwestern University, EvanstonIL 60208 USA (e-mail: jiazhang.wang@u.northwestern.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2021.3081873.

Digital Object Identifier 10.1109/TMM.2021.3081873

triplet architecture is adopted here, and triplet loss is proposed to constrain the irrelevant features and learn a highly discriminate and robust embedding related only to the head pose.

The second problem that needs to be addressed is how to represent the angles of the head pose effectively. In previous studies [7], [9], two representations, namely, Euler angles and unit quaternion, were often adopted. Using Euler angles can cause the ambiguity problem called gimbal lock [10]. If it occurs, the representation system will lose one degree of freedom, and the head poses will not be independent of one other and will rotate in all three directions simultaneously. Another representation is unit quaternion, which also suffers from the ambiguity problem since its double embedding may lead to the existence of two disconnected local minimal values. Moreover, Zhou *et al.* [11] demonstrated that any rotation representation in four or fewer dimensions is discontinuous, which is a nuisance in the optimization of the training network [12]. To overcome the ambiguity problem, the rotation matrix representation is adopted in this study. In geometry, the 3D rotation group, often denoted $\mathrm{SO}(3)$, is the group of all rotations around the origin of three-dimensional Euclidean space. The task of image-based HPE represented by rotation matrix is equivalent to estimating rotation matrix on $\mathrm{SO}(3)$. However, due to the uncertainty of the head pose, the rotation matrix is difficult to be predicted directly. For dealing with the pose uncertainty, various probability distributions for rotations are defined in directional statistics [13], [14]. The matrix Fisher distribution is an exponential probability density for rotation matrices introduced in [15], [16]. The matrix Fisher distribution on $\mathrm{SO}(3)$ can be defined by 9 parameters, and it is corresponding to the Gaussian distribution in $\mathbb{R}^3$ which is defined by the three dimensional mean, and the six dimensional covariance. The matrix Fisher distribution on $\mathrm{SO}(3)$ can subsequently constrain the rotation matrix and model pose uncertainty for further analysis and estimation.

The matrix Fisher distribution network (MFDNet) model is proposed based on these two solution schemes. MFDNet consists of triplet module and matrix Fisher distribution module. The main contributions of this work are summarized below.

1) Unlike traditional paradigms that learn embeddings from a single image, our method considers that the multiple images can provide additional information on the head pose. A triplet module employs triplet architecture with triplet loss is proposed, and it needs three images as the input, two of the images share the same pose, and the other image has another pose. Using the triplet loss, the difference in their embeddings is computed to obtain the highly discriminate and robust embeddings related to the head pose. The extracted embeddings can further enhance the pose tolerance.
2) In contrast to previous rotation representations, the matrix Fisher distribution based on the rotation matrix is utilized to represent the angles of the head pose and model the uncertainty of head rotation. The matrix Fisher distribution can effectively avoid the ambiguity problem and dramatically improve the accuracy of HPE.
3) Our method is evaluated on two protocols trained on 300W-LP and BIWI datasets. The experimental results demonstrate that the proposed method outperforms state-of-the-art methods based on RGB images and is even favorable against heavy methods adopting depth and time information on AFLW2000 and BIWI datasets.

The rest of the paper is structured as follows. Current work related to HPE is presented in Section II. The details of the MFDNet model and its different modules (triplet module and matrix Fisher distribution module) are introduced in Section III. The experimental results and discussion of different datasets are reported in Section IV. Section V concludes this study.

II. Related Work

A. Head Pose Estimation

Head pose estimation has existed as a studied task in computer vision in the past few years. Recently, with the rapid development of CNN, it is leveraged more commonly in estimating head pose. Many methods have been proposed to estimate head poses on different modalities of images, such as RGB images, depth images, and videos. The existing approaches for HPE on RGB images can be coarsely categorized into landmark-based and landmark-free methods.

Landmark-based methods predict head poses by solving the correspondence between 2D facial landmarks and 3D head models, therefore, they require an accurate facial landmark detector and a head model generator. Sun *et al.* [17] proposed a cascade convolution network to generate the facial landmark detector. In [18], [19], a supervision-by-registration approach was proposed to relieve the effects of inaccurate labels and improve the performance of facial landmark detection. A face alignment network (FAN) [5] that can extract facial landmarks through heatmap regression has also been presented. In [20], the spherical parameterization method was developed to generate a morphable 3D head model for determining the accurate projection relationship. Landmark-free methods directly estimate head poses without facial landmarks. These approaches mainly adopted Euler angles to represent the angles of head pose. A multi-task strategy that can localize facial landmarks and estimate head poses was adopted in [21], [22]. Zhu *et al.* [23] proposed a 3D dense face alignment (3DDFA) method that can utilize facial landmarks to fit a 3D face model and homeopathically estimate the head pose. HopeNet was presented in [7], which adopted a coarse-to-fine strategy by using classification and regression losses for constraining head poses. An improved method based on the HopeNet was presented in [24]. In this method, the direct regression procedure is modified to average top-k regression for estimation. Moreover, Yang *et al.* [8] exploited the soft stage-wise architecture and feature aggregation to obtain head poses. In [25], a spatial channel-aware residual attention structure (SCR-AT) was developed to construct Gaussian distribution as the ground truth and incorporate the attention module to estimate the head pose. A feature decoupling network (FAN) that can replace the Kullback Leibler (KL) divergence loss with cross entropy loss and calculate the intra-class gap of each Euler angle by using cross-category center loss was presented for HPE in [26]. In addition, unit quaternion has also been utilized to represent head poses. Hsu *et al.* [9] proposed a

quaternion network (QuatNet) to predict the angles of head pose represented by unit quaternion instead of Euler angle.

Several methods have also been developed for different image modalities, such as depth images and videos. Fanelli *et al.* [27] employed the random forest algorithm to estimate the head pose from depth images. In [28], DeepHeadPose that combined RGB images and depth information to regress head poses was presented. Meyer *et al.* [29] proposed a morphable face model and registered it to depth images for HPE. Moreover, in [30], Gu *et al.* utilized the recurrent neural network to solve facial videos for estimating head pose. Different from these approaches, the current study employs a triplet architecture with triplet loss. Then a matrix Fisher distribution based on the rotation matrix is proposed to represent the angles of the head pose for HPE on RGB images.

B. Metric Learning

Metric learning refers to the embedding of data with highly discriminate and robust capability based on a similarity measure defined by an optimal distance metric [31]. Metric learning is widely used for face recognition [32] and person reidentification [33], [34]. The Siamese network [35] and Triplet network [36], which are the twin or triplet architecture with the same subnetworks, respectively, are commonly leveraged methods. Early methods based on the Siamese network with contrastive loss [37] exhibit superior performance by measuring the distance of pairwise samples. Schroff *et al.* [32] proposed an effective paradigm and used the triplet network with triplet loss to minimize the distance between the positive pairs that share the same label while maximizing the distance between the negative pairs that have different labels. Moreover, Hermans *et al.* [38] claimed that the triplet loss and its variants outperform other published methods by a large margin. In view of the observed phenomenon regarding low pose tolerance, we argue that using multiple images as the input may provide additional information about head pose. Hence, the triplet network with triplet loss is adopted here to learn a discriminate and robust embedding related to head pose.

C. Rotation Matrix Representation

It is extremely important to determine the representation of the pose in pose estimation. Several studies that used deep learning model regarded the Euler angle [20] or unit quaternion [9] as representations for pose estimation. However, these representations suffer from the ambiguity problem due to their properties. Therefore, the rotation matrix is used for pose estimation to address this issue. Prokudin *et al.* [39] explored the rotation matrix combined with singular value decomposition for pose estimation. In [40], Lee adopted matrix Fisher distribution [15] to represent the uncertainties of attitude estimates and constructed a pose estimator according to the Bayesian framework. Mohlin *et al.* [41] proposed an approximating the normalizing constant of the matrix Fisher distribution method to estimate object orientation. Wang *et al.* [42] represented large uncertainties in pose by combining the matrix Fisher distribution with the Gaussian distribution. With regard to the representation problem commonly encountered in HPE, rotation matrix representation and matrix Fisher distribution not only avoid the ambiguity problem but also further analyze the uncertainty of head rotation.

III. Methodology

In this section, we firstly introduce the head pose estimation problem and describe some details of the proposed MFDNet model. Secondly, the triplet module for extracting a more discriminate and robust embedding related to head pose is presented. Then, we introduce the rotation matrix representations and the matrix Fisher distribution module for predicting head pose. Finally, we present the optimization for training.

A. Problem Formulation

Generally, head pose estimation task can be briefly summarized as follows. Given a facial image x and its corresponding head pose y is represented by the Euler angles including yaw angle ϕ, pitch angle θ and roll angle ψ. HPE is to find a mapping function F to estimate $\hat{y} = F(x)$. The $\hat{y}$ should fit the real head pose y as much as possible. The function F can be obtained by minimizing the error of the prediction $\hat{y}$ and the ground truth y. The error is expressed as $L(\hat{y}, y)$ where is the loss function.

B. Overview of the Proposed MFDNet Model

The proposed MFDNet model is shown in Fig. 2, which contains two modules, such as triplet module and matrix Fisher distribution module. During training phase, the model can be trained with an end-to-end strategy. It employs triplet network architecture, which takes three images as input and fed them into the backbone to obtain embeddings related to head pose. These extracted embeddings are firstly put into triplet module to compute triplet loss for improving the robustness of the network. Then these feature maps are passed to fully connected layers to obtain the unconstrained matrix M for constructing matrix Fisher distribution, and calculate MFD loss with ground truth respectively. In the end, we combinate the triplet loss and three MFD losses to train model. In the inference phase, the triplet module will be discarded and MFDNet takes one image as input. Then the extracted embedding is entered into matrix Fisher distribution module to generate the predicted head pose represented by rotation matrix. In addition, considering the network depth and width, EfficientNet-b0 [43] is selected as the feature extractor in our model.

The details of two modules and the optimization will be described in subsequent sections.

C. Robust Pose Embedding Extraction With Triplet Module

Unlike the previous methods for head pose estimation, we take multiple images as inputs for obtaining additional information about head pose. The selection of triplet samples depends on the different coarseness of head pose. Considering the time cost and computation efficiency, we firstly define the binned poses which are partitioned into 12 bins separated by 15 degrees. Then triplet samples (x^n, x^a, x^p) are randomly selected as inputs according to binned pose, where (x^n, x^a) indicates different binned poses, and (x^a, x^p) indicates the same binned

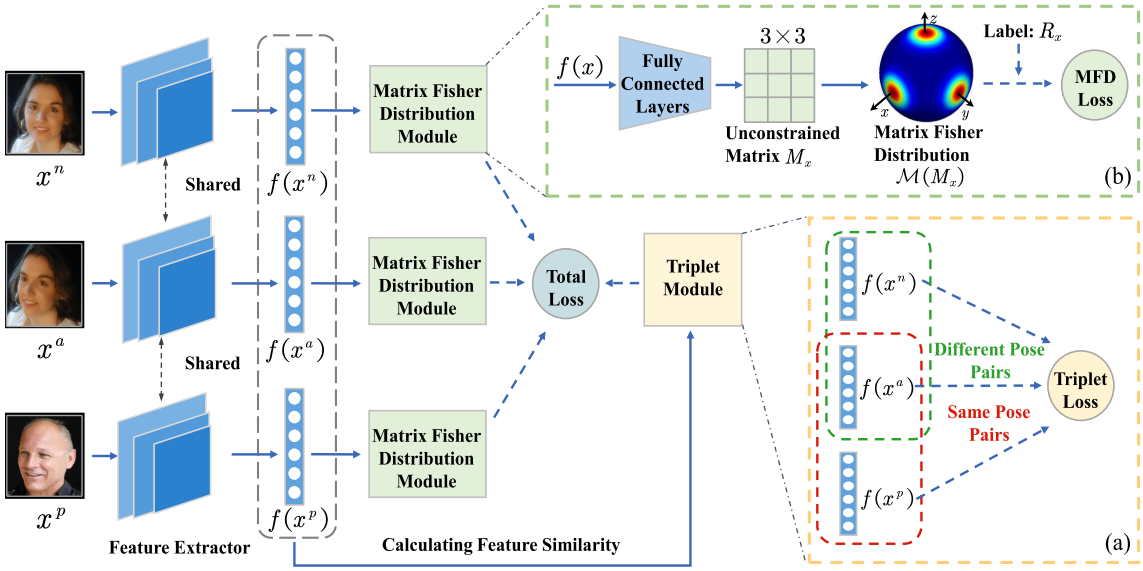


Fig. 2. Illustration of the proposed MFDNet model. (a) The triplet module that computes the difference of three embeddings. (b) The matrix Fisher distribution module that predict head pose.

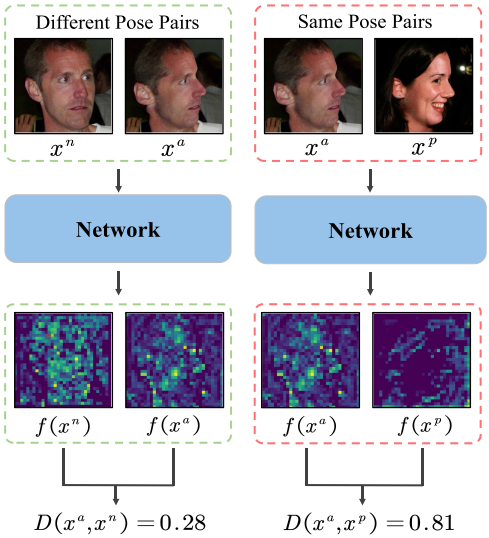


Fig. 3. Comparison of the distance between embeddings of different image pairs. The first row shows the different pose pairs and the same pose pairs, the second row is the feature extractor using HopeNet. The third row displays the extracted embeddings. Lastly the distance of embeddings from different poses and the same pose pairs is computed.

pose. The triplet samples are input into the backbone to obtain the embeddings $f(x^n)$, $f(x^a)$ and $f(x^p)$ respectively. Then we take advantage of Euclidean distance $D(\cdot,\cdot)$ to measure the similarity between different embeddings. This measurement can be expressed as

$$D(x_1, x_2) = \| f(x_1) - f(x_2) \|_2^2. \tag{1}$$

Considering the low pose tolerance, as illustrated in Fig. 3, HopeNet [7] is utilized to extract the embeddings of triplet samples. By computing their Euclidean distance, it can be observed that the similarity of embeddings with different pose pairs is smaller than the distance of embeddings with the same pose pairs.

However, we expect $f(x^a)$ to be more similar to $f(x^p)$ than $f(x^n)$, which can be expressed by $D(x^a, x^n) < D(x^a, x^p)$. Therefore, the triplet module is proposed as illustrated in Fig. 2(a), where the green dashed box and red dashed box separately indicate the embeddings from different pose pairs and the same pose pairs. The triplet loss is utilized to maximize the distance between embeddings with different pose pairs and minimize the distance between embeddings with the same pose pairs, which can obtain more robust embeddings related only to head pose. To ensure that the network extracts enough variation between embeddings with different poses, the class margin γ is used to control this similarity variation. Thus, the triplet loss is defined as

$$\mathcal{L}_{triplet}(x^n, x^a, x^p) = max(0, D(x^a, x^n) - D(x^a, x^p) + \gamma). \tag{2}$$

D. Modeling the Pose Uncertainty With Matrix Fisher Distribution

In the section, we firstly proposed rotation matrix as representation for head pose. Then matrix Fisher distribution based on rotation matrix will be introduced to model the head pose rotation uncertainty, while computation methods are discussed for its normalizing constants. Finally, the negative log-likelihood loss is proposed to measure the validity of the distribution.

1) Head Pose Representations: The assumptions that take rotation matrix as the representation for the angles of head pose is introduced. In the 3D world coordinate, the point located at (u, v, w) firstly rotates angle ψ around x-axis, then rotates angle θ around y-axis and finally rotates angle ϕ around z-axis. Such a sequence of rotations can be expressed by

$$R_z(\phi) R_y(\theta) R_x(\psi) \cdot (u, v, w)^T = R \cdot (u, v, w)^T, \tag{3}$$

where R is the rotation matrix, which can represent the linear transformation from the body-fixed frame to the inertial frame. The angles of head pose can be represented by rotation matrix R. It can be obtained as (4) shown at the bottom of this page.

where the angles ϕ, θ and ψ are the Euler angles and indicate yaw, pitch and roll angles, respectively. In addition, for the existing head pose estimation datasets, the labels are represented by Euler angles. These labels should be transformed into rotation matrix. The Euler angles can be determined by the element of the rotation matrix. The transformation can be expressed by

$$\begin{cases} \psi = \text{atan2}\left(R_{32}, R_{33}\right), \\ \theta = \text{atan2}\left(-R_{31}, \sqrt{R_{32}^2 + R_{33}^2}\right), \\ \phi = \text{atan2}\left(R_{21}, R_{11}\right). \end{cases} \tag{5}$$

The 3D rotation group, often denoted SO(3), is the group of all rotations around the origin of three-dimensional Euclidean space. Rotation matrix is the three-dimensional special orthogonal matrix on Lie group SO(3) and satises the following properties

$$\text{SO}(3) = \left\{R \in \mathbb{R}^{3\times 3} \mid R^T R = I_{3\times 3}, det\left[R\right] = +1\right\}, \tag{6}$$

where $det[\cdot]$ is the determinant of a square matrix.

Head pose estimation can be designed on the three dimensional special orthogonal group to escape the singularity and ambiguity problem. Hence, the rotation matrix is difficult to constrain, so it is easy to lose its orthogonality. Therefore, in view of constructing deterministic head pose on SO(3), we construct a compact form of an exponential density model on the special orthogonal group, namely matrix Fisher distribution, to constrain rotation matrix and represent head pose rotation uncertainty.

2) Matrix Fisher Distribution: The matrix Fisher distribution introduced in [15], [16] deals with statistics for rotations. When applied to the three-dimensional special orthogonal group, it can be determined by 9 parameters and modeling the head pose uncertainty. For a rotation matrix $R \in \text{SO}(3)$, the matrix Fisher distribution is defined by the following probability density function

$$p\left(R \mid M\right) = \frac{1}{a(M)} \exp\left(tr\left[M^T R\right]\right), \tag{7}$$

where $M \in \mathbb{R}^{3\times 3}$ is an unconstrained matrix, $a(M) \in \mathbb{R}$ is the normalizing constant of distribution and $tr[\cdot]$ is the trace of the matrix. The distribution can be denoted by $R \sim \mathcal{M}(M)$.

The normalizing constant $a(M)$ is given by

$$a(M) = \int_{R\in SO(3)} \exp\left(tr\left[M^T R\right]\right) dR, \tag{8}$$

which is rather non-trivial to compute accurately and efficiently. To evaluate $a(M)$, the singular values of M are considered. Assuming the singular value decomposition of unconstrained matrix M is given by

$$M = USV^T, \tag{9}$$

where $S = diag(s_1, s_2, s_3)$ is a diagonal matrix of singular values of M and $s_1 \geq s_2 \geq s_3 \geq 0$. The U and V are the orthogonal matrices which have $U^T U = V^T V = I_{3\times 3}$. While they are not rotation matrices in SO(3) since the determinant of U and V could be -1. To address this issue, we apply this transformation as follow

$$\begin{cases} \hat{U} = U diag\left(1, 1, det\left[U\right]\right), \\ \hat{S} = diag\left(s_1, s_2, s_3'\right) = diag\left(s_1, s_2, det\left[UV\right] s_3\right), \\ \hat{V} = V diag\left(1, 1, det\left[V\right]\right). \end{cases} \tag{10}$$

The definition of $\hat{U}$ and $\hat{V}$ ensure $\hat{U}, \hat{V} \in \text{SO}(3)$. Moreover, the result of $a(M)$ is determined by $\hat{S}$. The Λ is defined as

$$\Lambda = diag(s_1 - s_2 - s_3', s_2 - s_1 - s_3', s_3' - s_1 - s_2, s_1 + s_2 + s_3'), \tag{11}$$

the normalizing constant $a(M)$ can be expressed by

$$a(M) = {}_1F_1\left(\frac{1}{2}, 2, \Lambda\right), \tag{12}$$

where ${}_1F_1(\cdot,\cdot,\cdot)$ is the generalized hypergeometric function, which is rather non-trivial to compute accurately. By computing this hypergeometric function, the constant $a(M)$ and its gradients can be obtained.

When acquired the matrix Fisher distribution on SO(3), in the inference phase, the proper rotation matrix can be extracted from the distribution. Suppose $R \in \text{SO}(3)$ and $R \sim \mathcal{M}(M)$, the maximum mean can be defined as

$$R_{\text{MM}} = \underset{R\in\text{SO}(3)}{\text{argmax}}\{p(R|M)\}. \tag{13}$$

After calculating, the optimal rotation $\hat{R} \in \text{SO}(3)$ is expressed by

$$\hat{R} = R_{\text{MM}} = \hat{U}\hat{V}^T = U \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & det\left[UV\right] \end{bmatrix} V^T. \tag{14}$$

In order to directly visualize the matrix Fisher distribution, a method for visualizing the probability density function on SO(3) was proposed. We firstly compute the marginal probability density function for the *i*-th column vector of rotation matrix, and visualize it on the surface of the unit sphere with color map. The matrix Fisher distribution is shown in Fig. 4. The matrix M determines the shape of $\mathcal{M}(M)$. The larger the singular value of M, the smaller the dispersion of the marginal probability density. Moreover, as shown in Fig. 4(d), the principal axes shown by the red axes are rotated when M varies.

3) Negative Log-Likelihood Loss Function: Given a sample (x, R_x), x is the input image and the rotation matrix $R_x \in \text{SO}(3)$ is ground truth. The Fig. 2(b) shows the details of matrix Fisher

$$R = \begin{bmatrix} \cos\theta\cos\phi & \sin\psi\sin\theta\cos\phi - \cos\psi\sin\phi & \cos\psi\sin\theta\cos\phi + \sin\psi\sin\phi \\ \cos\theta\sin\phi & \sin\psi\sin\theta\sin\phi + \cos\psi\cos\phi & \cos\psi\sin\theta\sin\phi - \sin\psi\cos\phi \\ -\sin\theta & \sin\psi\cos\theta & \cos\psi\cos\theta \end{bmatrix}, \tag{4}$$

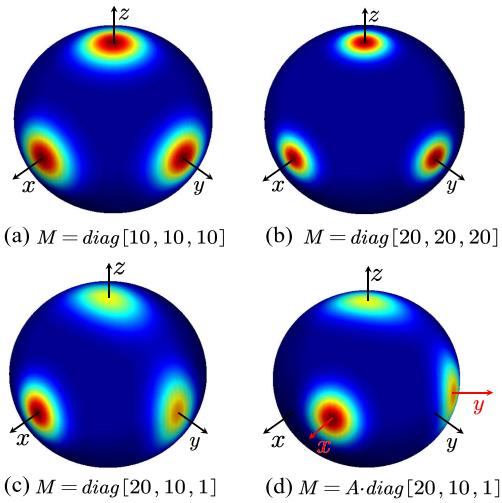


Fig. 4. Visualization of the matrix Fisher distributions with different unconstrained matrix M. For (a) and (b), the distributions on the three axes are similar and circular, but the distribution in (a) is more dispersive than in (b) due to the smaller singular values of M. (c) The difference in the singular values of M results in a concentrated distribution on the x-axis and the elongated distributions on the other axes. In (d), A is the rotation matrix given by rotating around the z-axis by $\pi/6$ degrees. By left-multiplying the matrix M with A, both the x-axis and y-axis are rotated but the shape of distribution remains as in (c).

Algorithm 1: Training Procedure For The Matrix Fisher Distribution Network.

Input: Training set T and the hyper parameters α, β.
Output:<**year**> Network parameters θ.
1: Initialize the parameters θ;
2: **while** not convergence **do**
3: Sampling triplet samples $X=(x^n, x^a, x^p)$ from T;
4: Generating rotation matrix label $Y=(R_{x^n}, R_{x^a}, R_{x^p})$;
5: Computing $\mathcal{L}_{triplet}$ with X according to (2);
6: Calculating $\mathcal{L}_{MFD}$ with (X, Y) according to (15);
7: $\theta \leftarrow \theta - \nabla_\theta(\alpha\mathcal{L}_{triplet}(X) + \beta\mathcal{L}_{MFD}(X, Y))$;
8: **return**θ.

distribution module. By feeding the sample to the convolutional neural network and fully connected layers sequentially, we obtain the 9D representations M_x, then generate matrix Fisher distribution $\mathcal{M}(M_x)$. The matrix Fisher distribution loss (MFD loss) between M_x and R_x can be computed by utilizing the negative log-likelihood function of distribution. The MFD loss can be expressed by

$$\mathcal{L}_{MFD}(x) = -\log(p(R_x|M_x)) = \log(a(M_x)) - tr\left[M_x^T R_x\right]. \quad (15)$$

This loss is a continuous convex function and has continuous gradients which makes it suitable for optimization.

E. Optimization

The proposed MFDNet model is trained by minimizing the total loss as shown in Fig. 2. Regarding to the losses at different modules, the total loss $\mathcal{L}$ consists of the above losses and can be formulated as

$$\mathcal{L}(x^n, x^a, x^p) = \alpha\mathcal{L}_{triplet}(x^n, x^a, x^p) + \beta[\mathcal{L}_{MFD}(x^n) + \mathcal{L}_{MFD}(x^a) + \mathcal{L}_{MFD}(x^p)] \quad (16)$$

where α, β are the hyperparameters that can balance the weights of different losses. The MFDNet model can be supervised by $\mathcal{L}$. The training procedure of the proposed method is summarized in Algorithm 1.

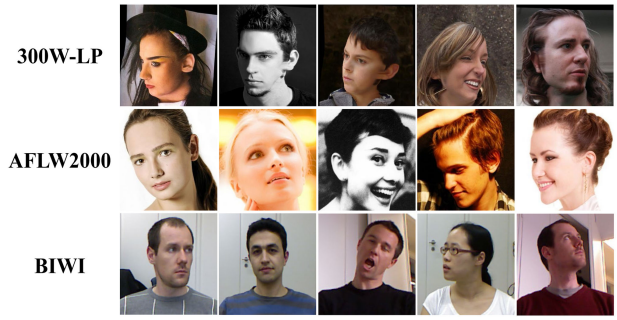


Fig. 5. HPE images in three datasets. The first, second and third rows are samples from 300W-LP, AFLW2000 and BIWI datasets, respectively.

IV. Experiment

A. Experiment Settings

1) Datasets: In this experiment, three public head pose datasets are used. Some samples of datasets can be shown in Fig. 5.

300W-LP dataset [23] contains more than 120 K images across large poses by meshing and flipping 300 W dataset [44] which provides annotations for 3837 face images with 68 landmarks.

BIWI dataset [45] records the different head pose rotations by Kinect v2 device and contains over 24 videos of 20 subject, a total of 15 K frames. For each frame, both RGB images and depth images are provided.

AFLW2000 dataset [23] contains the first 2000 images of the AFLW [46] whose annotations include large head poses and 68 landmarks.

2) Evaluation Protocols: For training and evaluating the proposed methods on these different datasets, our experiments are executed according the following two widely used protocols same as FSA-Net [8].

In protocol I, the 300W-LP dataset is used for training, and the trained model is evaluated on two real-world datasets, the AFLW2000 and BIWI datasets. When testing on AFLW2000, we cropped around the head region based on the facial landmark labels. When testing on BIWI, we employ MTCNN face detector [47] for face detection instead of using tracking.

In protocol II, the 70% videos in the BIWI dataset are used as training sets and the others for testing. The videos are also cropped head region using MTCNN face detector. Notice that, the videos include different modalities like RGB, depth and time. The proposed method only works with RGB frames.

For a fair comparison with other methods, we need to transform the predicted rotation matrix into Euler angles for evaluation. Given an image x_i, we can predict its rotation matrix R_i using MFDNet. Then the Euler angle can be transformed according to (5). The mean absolute error (MAE) is adopted as the evaluation metric, it is defined as follows

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|\hat{y}_i - y_i|, \tag{17}$$

where m is the number of samples on the datasets, $\hat{y}_i$ and y_i denote the predicted Euler angle and the ground truth of the *i*-th sample respectively.

3) Implementation: The experiments were carried out based on PyTorch and performed on a workstation with Intel Xeon Gold 6126 CPU and Nvidia Tesla V100 GPU. All the input images are firstly normalized using the ImageNet [48] mean and standard deviation, then randomly cropped to size 224×224. Regarding the hyperparameters on the model, the class margin γ is set to 0.2 in the triplet module same as [36]. The weights of the different losses in the total loss are $\alpha = 2$ and $\beta = 3$. Adam optimizer is adopted to update the model parameters for 30 epochs and the batch size is 32. The learning rate is initialized to 1e-4 and reduced by a factor of 0.1 every 10 epochs.

B. Experiment Results Analysis

1) Competing Methods: To demonstrated the effectiveness of the proposed methods, we compare our methods with several state-of-the-art methods.

Plenty of methods on RGB images are compared. Dlib [6] and FAN [5] solved the 2D to 3D fitting problem via facial landmarks for head pose estimation. KEPLER [21] utilized multi-tasking strategy to predict both landmark and head pose simultaneously. 3DDFA [23] fitted a 3D face model to RGB image and generated head pose. The following methods use landmark-free scheme to predict. HopeNet [7] employed fine-grained strategy to build a multi-loss for Euler angles. Moreover, for creating a fair comparison, we constructed a new benchmark based on HopeNet, which called HopeNet+. It adopts EfficientNet-b0 instead of ResNet-50 as backbone same as our proposed method. FSA-Net [8] proposed a soft stage-wise regression method and combined with feature aggregation to achieve good performance. QuatNet [9] taken advantage of uint quaternion to represent head pose. SCR-AT [25] combined with Gaussian distribution and spatial attention structure to predict Euler angles. HPE-40 [24] proposed a novel head pose estimation method using two-stage ensembles with top-*k* regression. FDN [26] built a feature decoupling network to learn exclusive features of different angles for head pose estimation.

Additional methods on different modalities also is compared. DeepHeadPose [28] adopted course-to-fine strategy to estimate the head pose from RGB-D images. Gu *et al* [30] employed VGG16 and VGG16 + RNN to predict to predict head pose for RGB images and RGB + Time images by combing Bayesian filters. Martin [49] constructed a 3D head model to estimate the head pose from depth images.

TABLE I
COMPARISON TO OTHER STATE-OF-THE-ART METHODS ON AFLW2000 DATASET. ALL METHODS ARE TRAINED ON 300W-LP DATASET

Methods	Yaw	Pitch	Roll	MAE
Dlib [6]	23.1	13.6	10.5	15.8
3DDFA [23]	5.40	8.53	8.25	7.39
FAN [5]	6.36	12.3	8.71	9.12
HopeNet [7]	6.47	6.56	5.44	6.16
HopeNet+	5.57	6.24	5.02	5.61
FSA-Net [8]	4.50	6.08	4.64	5.07
QuatNet [9]	3.97	5.62	3.92	4.50
SCR-AT + DGDL [25]	**3.77**	5.35	4.06	4.39
HPE-40 [24]	4.87	6.18	4.80	5.28
FDN [26]	3.78	5.61	3.88	4.42
MFDNet($\alpha = 2$, $\beta = 3$)	4.30	**5.16**	**3.69**	**4.38**

TABLE II
COMPARISON TO OTHER STATE-OF-THE-ART METHODS ON BIWI DATASET. ALL METHODS ARE TRAINED ON 300W-LP DATASET

Methods	Yaw	Pitch	Roll	MAE
Dlib [6]	16.8	13.8	6.19	12.2
3DDFA [23]	36.2	12.3	8.78	19.1
FAN [5]	8.53	7.48	7.63	7.89
KEPLER [21]	8.80	17.3	16.2	13.9
HopeNet [7]	5.17	6.98	3.39	5.18
HopeNet+	4.85	5.96	3.06	4.62
FSA-Net [8]	4.27	4.96	2.76	4.00
QuatNet [9]	4.01	5.49	2.94	4.15
SCR-AT + DGDL [25]	3.63	**4.46**	3.08	3.72
HPE-40 [24]	4.57	5.18	3.12	4.29
FDN [26]	4.52	4.70	**2.56**	3.93
MFDNet($\alpha = 2$, $\beta = 3$)	**3.40**	4.68	2.77	**3.62**

2) Results With Protocol I: In this scenario, our model is trained on 300W-LP dataset then tested on AFLW2000 and BIWI datasets. Compared with the above methods, the performance of our method is better than theirs. Table I demonstrates the result evaluated on AFLW2000 dataset. The proposed MFDNet always reflects the lowest variance on the errors of pitch and roll angles. When the weights of triplet loss and matrix Fisher distribution loss are $\alpha = 2$ and $\beta = 3$, the MFDNet achieves the lowest MAE and has reduced the MAE to 4.38. In fact, our method is designed specifically for the head pose estimation task. It can mitigate the low pose tolerance and provide more robust and accurate performance.

A similar situation that the proposed approach is outstanding on BIWI dataset is reported in Table II. Since BIWI dataset was collected in an experimental environment, its low disturbance leads to lower MAE than the results on other dataset. The proposed MFDNet employs matrix Fisher distribution to model the uncertainty of head rotation. It is able to further constrain the rotation matrix and more accurately predict the corresponding pose. Hence, the MFDNet where $\alpha = 2$ and $\beta = 3$ performs the lowest error on predicting yaw angle and achieves the state-of-the-art with a smallest MAE.

3) Results With Protocol II: The BIWI dataset is used for training and testing respectively. As Table III shown, the experimental results for different modalities of images on BIWI dataset are displayed. The RGB groups only use RGB images,

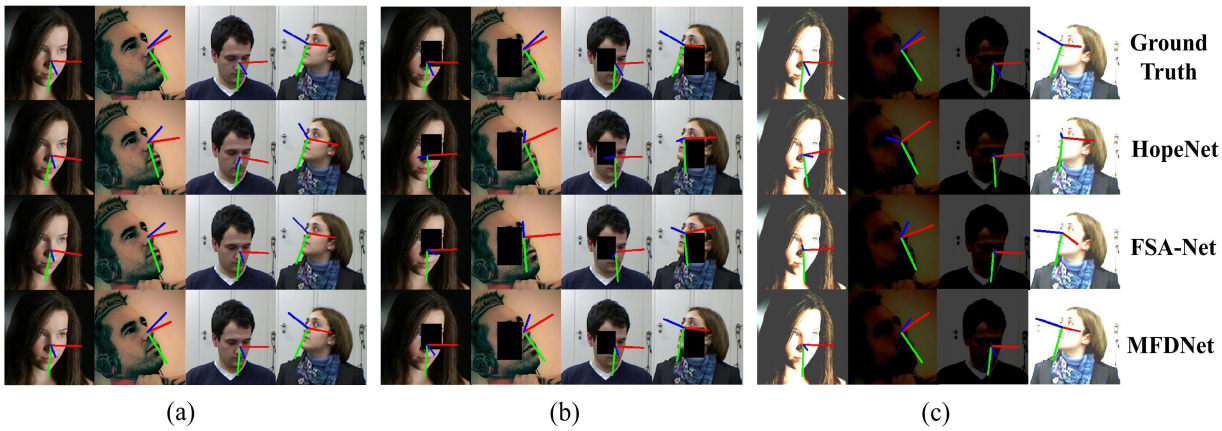


Fig. 6. Head pose estimation on the samples from AFLW2000 and BIWI datasets. These samples including the original images (a), the corresponding occluded images (b) and the corresponding images in different illumination conditions (c). From top to bottom, they are ground truth, the predictions by HopeNet, the results of FSA-Net and our results. The blue, green and red lines indicate the front, downward and side of the face respectively.

TABLE III
COMPARISON TO OTHER STATE-OF-THE-ART METHODS ON BIWI DATASET. 70% OF BIWI DATASET ARE APPLIED FOR TRAINING AND THE OTHERS FOR TESTING. DIFFERENT TYPES OF METHODS LEVERAGE DIFFERENT MODALITIES OF DATASET, INCLUDING RGB, RGB + DEPTH AND RGB + TIME

Methods	Yaw	Pitch	Roll	MAE
RGB				
DeepHeadPose [28]	5.67	5.18	-	-
VGG16 [30]	3.91	4.03	3.03	3.66
FSA-Net [8]	**2.89**	4.29	3.60	3.60
FDN [26]	3.00	3.98	**2.88**	3.29
MFDNet($\alpha = 2$, $\beta = 3$)	2.99	**3.68**	2.99	**3.22**
RGB+Depth				
DeepHeadPose [28]	5.23	4.76	-	-
Martin [49]	3.60	2.50	2.60	2.90
RGB+Time				
VGG16 + RNN [30]	3.14	3.48	2.60	3.07

whereas RGB + Depth and RGB + Time groups make additional use of depth and temporal information. Our proposed MFDNet performs best among RGB-based methods, which reduces the MAE to 3.22 on BIWI test sets. Not only is it extremely close to the multimodality approach in terms of MAE, but it also achieves state-of-the-art on predicting yaw angle.

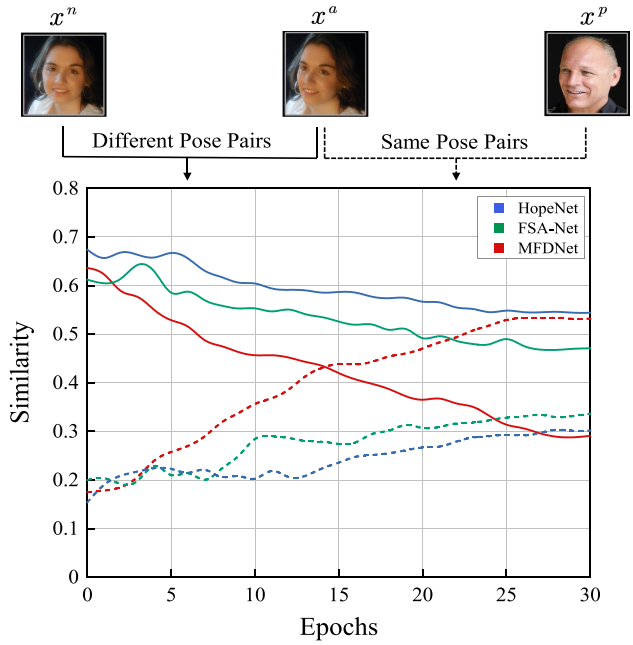


Fig. 7. Comparison of the variation in similarity between embeddings from different images extracted by different methods during training. The solid line shows the similarity between the embeddings from image pairs with different poses, while the dashed line is the embeddings from image pairs with the same pose. The red, blue, and green colors denote the change in similarity using MFDNet (our proposed), HopeNet, and FSA-Net methods.

C. Visualization

In this section, the performance compared with other methods and the training process are depicted. Firstly, the visualization of the performance of our approach is shown in Fig. 6. Randomly sampling samples from the AFLW2000 and BIWI datasets, we compare the performance of our proposed MFDNet model with HopeNet and FSA-Net. As Fig. 6(a) shown, the angles predicted by the MFDNet model is closer to the ground truth. Furthermore, some distortions including occlusion and illumination are applied to the samples as illustrated in Fig. 6(b) and Fig. 6(c). Different methods estimate head pose on these disturbed images for evaluating the robustness. It can be observed that our method not only has higher accuracy on estimating normal images, but also maintains a stronger robustness on estimating disturbed images. However, other methods lose the effect since the low pose tolerance.

To further illuminate the training process of our proposed method, the different modules will be visualized. The triplet module is specifically designed for extracting the embedding related to head pose. During training, as illustrated in Fig. 7, the similarity of embeddings from image pairs with different pose descends (solid line), the similarity of embeddings from image

TABLE IV
PERFORMANCE OF THE DIFFERENT BACKBONE STRUCTURES ON AFLW2000 AND BIWI DATASETS. ALL METHODS ARE TRAINED ON 300W-LP DATASET

Backbone	MB	AFLW2000				BIWI			
		Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
ResNet-50	98	4.39	5.76	4.62	4.92	3.50	4.68	2.97	3.71
ResNet-152	232	4.69	5.81	4.16	4.88	**3.37**	4.99	2.60	3.65
Inception-v4	184	4.64	5.46	3.86	4.65	3.72	**4.59**	2.96	3.76
DenseNet-201	80	4.85	5.79	4.21	4.95	3.62	5.72	**2.63**	3.99
EfficientNet-b0	**29**	**4.30**	**5.16**	**3.69**	**4.38**	3.40	4.68	2.77	**3.62**

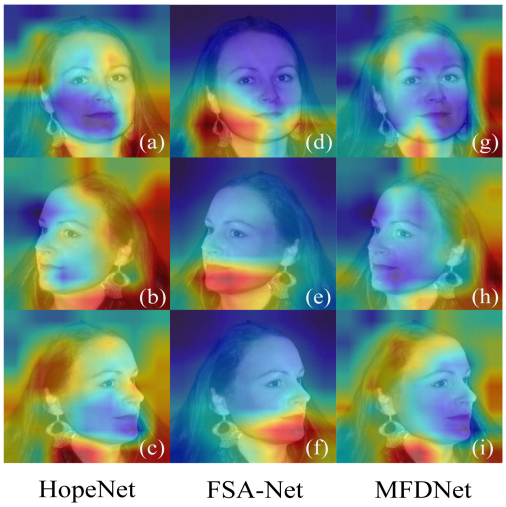


Fig. 8. Comparison of the generated Grad-CAM for the different methods on several samples with the same identity and different poses.

pairs with the same pose rises (dashed line) using our proposed MFDNet (red). The MFDNet makes the similarity of embeddings from images with the same pose progressively greater than the images with different poses. However, when using HopeNet (blue) and FSA-Net (green), the similarity curves change mildly, and the similarity of embeddings from images with different poses is still greater than images with the same pose until the end of training.

The Grad-CAM [50] can visualize the attended regions which contribute to head pose estimation. As Fig. 8 shown, the attended regions from different methods on the samples with same identity and different poses can be observed. The HopeNet and FSA-Net focus more on the facial regions, which may lead to wrong estimation because the samples with different poses share same facial regions. The proposed MFDNet mitigates the low pose tolerance using triplet module and eliminates the ambiguity problem caused by the same facial regions. It means that the MFDNet is able to extract the more robust and discriminate embeddings related to head pose.

Moreover, variations in the matrix Fisher distribution are visible during training. The Fig. 9(e) image is fed into MFDNet model after 10 epochs of training. The predicted matrix Fisher distribution for that image can be displayed as illustrated in Fig. 9(a-d). The peak of the marginal distribution of the rotation matrix is constantly evolving. The predicted distribution is constantly rotating towards the matrix Fisher distribution of ground truth, and its maximum likelihood estimation is constantly approaching the ground truth.

D. Ablation Study

In this section, the ablation studies are conducted to evaluate the performance of different backbone structures, analyze the parameters of the total losses and demonstrate the effectiveness of different components including the triplet module and the matrix Fisher distribution module. As Table IV shown, ResNet-50, ResNet-152, Inception-v4, DenseNet-201 and EfficientNet-b0 are selected as the backbone to extract the embeddings. All are trained on 300W-LP dataset then tested on AFLW2000 and BIWI datasets. The result exhibits the EfficientNet-b0 is not only a lightweight architecture but also offers better performance. Considering the tedious backbone selection, we can try to use Neural architecture search [51], [52] in the future work, and search the most suitable backbone for head pose estimation to further improve the performance based on the proposed method.

In addition, the triplet loss and MFD loss with different weights are summed to compose the total loss as in (16). The different weights in the total loss has an extremely important impact on the model. Therefore, we randomly divide the training sets and validation sets for the selection of weights. The 80% samples in the 300W-LP dataset are used for training and the others for validating. Meanwhile, in order to measure the generalization, we still evaluate the model using different weights on AFLW2000 dataset and BIWI dataset. The comparison of performance using different weights is presented as showed in Fig. 10. It can be observed that the MFDNet model achieves the best performance when the total loss is weighted $\alpha = 2$ and $\beta = 3$.

Furthermore, in order to verify the performance of the different components, we divide the proposed MFDNet model into three parts, w/o, Triplet Module and MFD Module. The first part w/o only adopts the feature extractor and fully connected layers without any proposed modules, which directly uses a single MSE to minimize the distance between the output unconstrained matrix M and ground truth R. The second method Triplet Module only adds triplet module based on the first method w/o architecture. The third method MFD Module also adopts the w/o architecture and only adds matrix Fisher distribution module, which constructs matrix Fisher distribution according to M then calculates the matrix Fisher distribution loss using the negative log-likelihood function. For all parts, the feature extractor is EfcientNet-b0. They are trained on 300W-LP dataset then tested on AFLW2000 and BIWI datasets. As Table V shown, the MFD Module method has smaller errors on both datasets, but it is less robust since the low pose tolerance. When fusing all modules together, as with MFDNet, it has stronger robustness and performs state-of-the-art results. Figure 11 illustrates

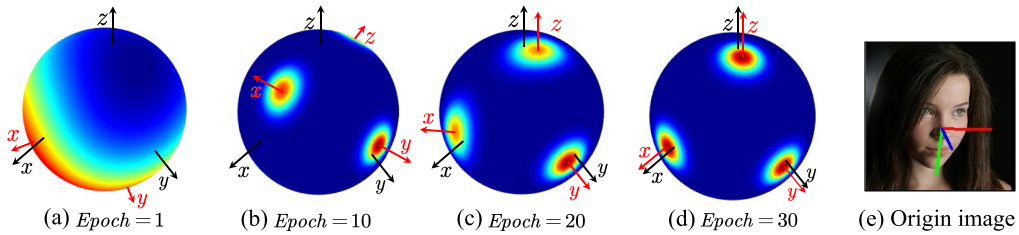


Fig. 9. Variation of the predicted matrix Fisher distribution during training. (a)-(d) indicate the matrix Fisher distribution of the head pose of image in (e) after the epoch of training, respectively, where the black axis displays the ground truth is represented by the rotation matrix and the red axis denotes the predicted rotation matrix obtained by computing the maximum likelihood estimation of the predicted distribution.

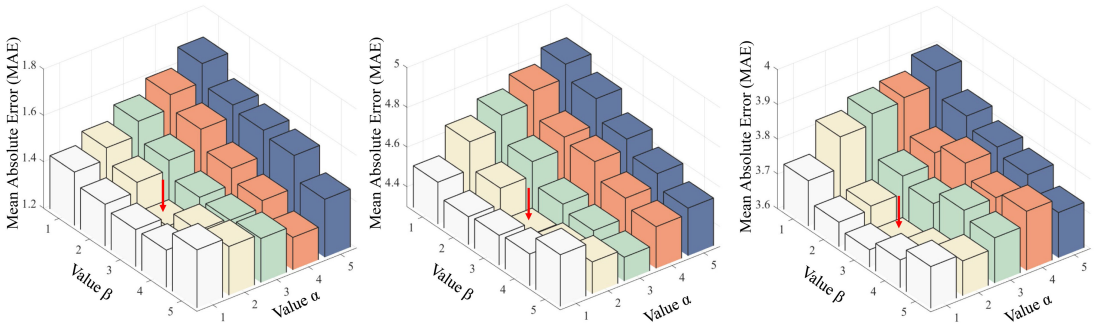


Fig. 10. Comparison of the MAE of MFDNet using total losses with different α and β. 80% of 300W-LP dataset are applied for training and the others for validating. All models are trained on 300W-LP train sets, then evaluated on 300W-LP validation sets, AFLW2000 and BIWI datasets respectively.

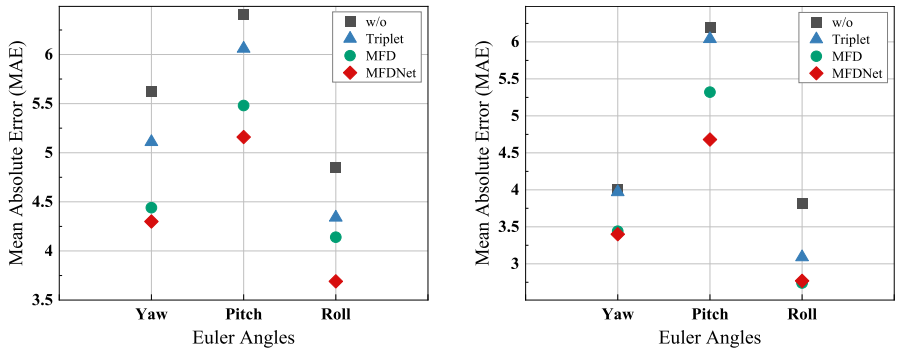


Fig. 11. Comparison of the MAE of different components at different angles under protocol I. MFDNet are divided into w/o, triplet module and matrix Fisher distribution module.

TABLE V
ABLATION STUDY OVER DIFFERENT COMPONENTS (W/O, TRIPLET MODULE, MFD MODULE AND MFDNET) ON AFLW2000 AND BIWI DATASETS. ALL ARE TRAINED ON 300W-LP DATASET

Methods	AFLW2000	BIWI
w/o	5.63	4.66
Triplet Module	5.17	4.36
MFD Module	4.69	3.83
MFDNet($\alpha = 2$, $\beta = 3$)	**4.38**	**3.62**

the details of the experimental results of the various modules at different angles.

V. CONCLUSION

In this work, a robust three-branch model called MFDNet, is proposed for estimating head poses from RGB images. MFDNet consists of triplet module and matrix Fisher distribution module. By designing the triplet module for three inputs, which are different pose pairs and the same pose pairs, the disturbance (including identity, illumination and occlusion) can be effectively limited, and embeddings that specifically represent the head pose can be obtained. As a result, robustness is considerably enhanced. In order to avoid the ambiguity problem when Euler angles and unit quaternion are used, the matrix Fisher distribution based on the rotation matrix is utilized to represent the angles of the head

pose, and model the uncertainty of head rotation for further constraining the rotation matrix and improving the performance of head pose estimation. Furthermore, the experimental results on two protocols indicate that the proposed MFDNet has stronger robustness than previous works and demonstrates state-of-the-art performance.

REFERENCES

[1] J. Lee and S. C. Marsella, "Predicting speaker head nods and the effects of affective information," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 552–562, Oct. 2010.

[2] S. Park, A. Spurr, and O. Hilliges, "Deep pictorial gaze estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 741–757.

[3] R. Stiefelhagen *et al.*, "Natural human-robot interaction using speech, head pose and gestures," in *Proc. Int. Conf. Intell. Robots Syst.*, 2004, pp. 2422–2427.

[4] K. Otsuka, "Behavioral analysis of kinetic telepresence for small symmetric group-to-group meetings," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1432–1447, Jun. 2018.

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

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

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

[8] T. Yang, Y. Chen, Y. Lin, and Y. Chuang, "FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1087–1096.

[9] H. Hsu, T. Wu, S. Wan, W. H. Wong, and C. Lee, "QuatNet: Quaternion-based head pose estimation with multiregression loss," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1035–1046, Apr. 2019.

[10] V. Lepetit and P. Fua, "Monocular model-based 3 d tracking of rigid objects: A survey," *Foundations Trends Comput. Graph. Vis.*, vol. 1, no. 1, pp. 1–89, 2005.

[11] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5745–5753.

[12] A. Saxena, J. Driemeyer, and A. Y. Ng, "Learning 3-d object orientation from images," in *Proc. Int. Conf. Robot. Automat.*, 2009, pp. 794–800.

[13] K. Mardia and P. Jupp, "*Directional Statistics*," Hoboken, NJ, USA: Wiley, 1999.

[14] Y. Chikuse, "*Statistics on Special Manifolds*." Berlin, Germany: Springer, 2003.

[15] T. Downs, "Orientation statistics," *Biometrika*, vol. 59, no. 3, pp. 665–676, 1972.

[16] C. Khatri and K. Mardia, "The von mises-fisher matrix distribution in orientation statistics," *Roy. Stat. Soc.*, vol. 39, no. 1, pp. 95–106, 1977.

[17] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3476–3483.

[18] X. Dong *et al.*, "Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 360–368.

[19] X. Dong *et al.*, "Supervision by registration and triangulation for landmark detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 1, p. 99, Mar. 2020.

[20] H. Yuan, M. Li, J. Hou, and J. Xiao, "Single image-based head pose estimation with spherical parametrization and 3D morphing," *Pattern Recognit.*, vol. 103, p. 107316, 2020.

[21] A. Kumar, A. Alavi, and R. Chellappa, "KEPLER: Keypoint and pose estimation of unconstrained faces by learning efficient H-CNN regressors," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 258–265.

[22] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.

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

[24] B. Huang, R. Chen, W. Xu, and Q. Zhou, "Improving head pose estimation using two-stage ensembles with top-k regression," *Image Vis. Comput.*, vol. 93, p. 103827, 2020.

[25] Y. Zhang, K. Fu, J. Wang, and P. Cheng, "Learning from discrete gaussian label distribution and spatial channel-aware residual attention for head pose estimation," *Neurocomputing*, vol. 407, pp. 259–269, 2020.

[26] H. Zhang, M. Wang, Y. Liu, and Y. Yuan, "FDN: Feature decoupling network for head pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12 789–12 796.

[27] G. Fanelli, T. Weise, J. Gall, and L. V. Gool, "Real time head pose estimation from consumer depth cameras," in *Proc. Joint Pattern Recognit. Symp.*, 2011, pp. 101–110.

[28] S. S. Mukherjee and N. M. Robertson, "Deep head pose: Gaze-direction estimation in multimodal video," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2094–2107, Nov. 2015.

[29] G. P. Meyer, S. Gupta, I. Frosio, D. Reddy, and J. Kautz, "Robust model-based 3 d head pose estimation," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 3649–3657.

[30] J. Gu, X. Yang, S. D. Mello, and J. Kautz, "Dynamic facial analysis: From bayesian filtering to recurrent neural network," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1531–1540.

[31] M. Kaya and H. S. Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, no. 9, p. 1066, 2019.

[32] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[33] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical clustering with hard-batch triplet loss for person re-identification," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13 654–13 662.

[34] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 3701–3711.

[35] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 539–546.

[36] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Similarity-Based Pattern Recognit.*, 2015, pp. 84–92.

[37] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1735–1742.

[38] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.

[39] S. Prokudin, P. V. Gehler, and S. Nowozin, "Deep directional statistics: Pose estimation with uncertainty quantification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 542–559.

[40] T. Lee, "Bayesian attitude estimation with the matrix fisher distribution on SO(3)," *IEEE Trans. Autom. Control*, vol. 63, no. 10, pp. 3377–3392, Oct. 2018.

[41] D. Mohlin, J. Sullivan, and G. Bianchi, "Probabilistic orientation estimation with matrix fisher distributions," in *Neural Inf. Process. Syst.*, pp. 4884–4893, 2020, *arXiv:2006.09740*.

[42] W. Wang and T. Lee, "Matrix fisher-Gaussian distribution on SO(3) $\times \mathbb{R}^n$ for attitude estimation with a gyro bias," in *Proc. Amer. Control Conf.*, 2020, pp. 4429–4434.

[43] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

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

[45] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. V. Gool, "Random forests for real time 3 d face analysis," *Int. J. Comput. Vis.*, vol. 101, no. 3, pp. 437–458, 2013.

[46] 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. Int. Conf. Comput. Vis. Workshops*, 2011, pp. 2144–2151.

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

[48] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

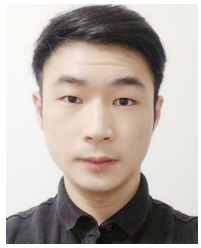
[49] M. Martinez, F. van de Camp, and R. Stiefelhagen, "Real time head model creation and head pose estimation on consumer depth cameras," in *Proc. Int. Conf. 3D Vis.*, 2014, pp. 641–648.

[50] R. R. Selvaraju *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
[51] C. Ying *et al.*, "Nas-bench-101: Towards reproducible neural architecture search," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7105–7114.
[52] X. Dong, L. Liu, K. Musial, and B. Gabrys, "NATS-Bench: Benchmarking nas algorithms for architecture topology and size," *IEEE Trans. Pattern Anal. Mach. Intell.*, Jan. 2021, doi: 10.1109/TPAMI.2021.3054824.

Hai Liu (Senior Member, IEEE) received the M.S. degree in applied mathematics and the Ph.D. degree in pattern recognition and artificial intelligence from the Huazhong University of Science and Technology, Wuhan, China, in 2010 and 2014, respectively. Since June 2017, he has been an Assistant Professor with the National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, China. His current research interests include deep learning, artificial intelligence, head pose estimation, gaze estimation, and pattern recognition.

Shuai Fang received the B.S. degrees from Sichuan University, Chengdu, China, in 2019. He is currently with the National Engineering Laboratory for Educational Big Data, Central China Normal University, Wuhan, china. His research interests include pattern recognition, head pose estimation, and human–machine interface.

Zhaoli Zhang received the M.S. degree in computer science from Central China Normal University, Wuhan, China, in 2004 and the Ph.D. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2008. Till February 2022, he is a Research Fellow with the City University of Hong Kong, Hong Kong. He is currently a Professor with the National Engineering Research Center for E-Learning, Central China Normal University. His research interests include computer vision and human–machine interface.



Duantengchuan Li is currently working toward the M.S. degree in computer science and technology with the National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, China. His research interests include representation learning, pattern recognition, head pose estimation, and their applications in industry and big data analysis.

Ke Lin is currently working toward the Ph.D. degree in control science and engineering with the Harbin Institute of Technology (Shenzhen), Shenzhen, China. His research interests include deep reinforcement learning, machine learning, and pattern recognition.

Jiazhang Wang is currently working toward the Ph.D. degree with Computational Photography Lab, ECE Department, Northwestern University, Evanston, IL, USA. His interests include pattern recognition, computational photography, deflectometry, and eye gazing estimation.