9D Rotation Representation-SVD Fusion with Deep Learning for Unconstrained Head Pose Estimation

Jiaqi Lyu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: lvjiaqi@st.xatu.edu.cn

Changyuan Wang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: Cyw901@163.com

***Abstract*—Accurately estimating human head pose poses a significant challenge across various application domains. To address the inherent limitations of previous approaches, this research proposes an unconstrained head pose estimation strategy. The method combines deep learning with rotation matrices, utilizing nine-dimensional vectors output by the neural network, which are projected back to rotation matrices in SO (3) space through singular value decomposition. This ensures both the smoothness and uniqueness of the rotation representation. The approach demonstrates distinct advantages in handling the rotation estimation task, particularly when the rotated representation is used as the model output. It not only avoids the discontinuity and double-coverage issues associated with prior methods but also enhances the stability of the representation in high-dimensional space, thereby improving the learning process. Additionally, the geodesic loss function is incorporated to train the network. The proposed strategy surpasses previous state-of-the-art methods, as evidenced by experiments conducted on the AFLW2000 and BIWI datasets.**

Keywords-Head Pose Estimation; Efficientnetv2; Rotation Matrix; Geodesic Loss

I. INTRODUCTION

In fields such as human-computer interaction [1] and augmented reality [2], head pose estimation has become a core technology driving immersive experiences and precise interactions. There are two main types of current methods: those that use landmarks and those that don't [3]. Landmark-based algorithms find important facial points in pictures and then use these points to map them to a 3D model of the head to figure out the 3D head position. Although this method is highly accurate, it is directly limited by the precision of key point localization. Occlusions and extreme rotation angles can make key points difficult to identify, leading to deviations in their positions and affecting the accuracy of the final head pose estimation.

Advancements in deep learning have significantly improved the accuracy of head pose estimation algorithms that do not depend on landmarks, because of the utilization of deep neural networks. HopeNet [4] proposes a multi-task learning approach that discretizes continuous head pose angles into several categories. It captures the discrete distribution of head poses through classification tasks while refining continuous angle values with regression tasks, using multi-task learning to predict Euler angles. QuatNet [5] employs a dual-branch structure for classification and regression. One branch uses a recurrent neural network for Euler angle classification, while the other represents head pose regression with quaternions. HPE [6] enhances head pose estimation by using a two-stage ensemble and a top-k regression. Multiple models independently predict in the first stage, and the top k optimal predictions are integrated in the second stage. WHENet [7] uses a single-branch model but increases the number of head pose angle categories. FSA-Net [8] utilizes a dual-branch architecture and fine-grained attention mechanism to effectively merge local and global image features, resulting in more accurate Euler angle predictions. TriNet [9] uses vectors to represent head direction instead of traditional Euler angles. FDN [10] introduces a feature decoupling method that helps the model focus on head pose-related features, ignoring background noise and other

DOI: 10.2478/ijanmc-2024-0028

irrelevant factors. LwPosr [11] is a lightweight network that employs a two-stream, three-stage structure for fine-grained regression. This structure combines a depth-separable convolution with a transformer encoder, enabling the network to efficiently predict head pose with a low number of parameters and high accuracy.

Many of the above methods split the rotation representation into bins for classification and combine it with regression for stable prediction, a practice that has become common. However, binning the angles can result in fragmented information. Additionally, choosing the appropriate rotation representation method is crucial for optimal performance. Most current methods use Euler angles or quaternions to train networks. While effective in some scenarios, they suffer from numerical discontinuities when handling large-scale and continuous rotations, such as the gimbal lock issue with Euler angles and the double coverage problem with quaternions. Zhou et al. [12] demonstrated that any rotation representation with four or fewer dimensions is discontinuous, making it unsuitable for neural network learning.

Geist et al. [13] summarized the characteristics of various rotation representations and their impact on gradient-based optimization methods. Building on this study, a head pose estimation technique is proposed that does not rely on landmarks but instead utilizes rotation matrices to accurately determine head pose direction. EfficientNetV2-S [14] is employed as the feature extraction network. Rather than directly predicting the rotation matrix, the neural network generates a nine-dimensional vector. This vector is subsequently transformed into a 3×3 matrix and converted into a valid rotation matrix using Singular Value Decomposition (SVD).

The network was trained using a geodesic loss function instead of the more commonly employed mean squared error (MSE) loss function. This choice was made because the geodesic loss function more effectively captures the differences in the manifold's rotations. The proposed approach is illustrated in Figure 1. The following sections provide a more detailed explanation of each component.

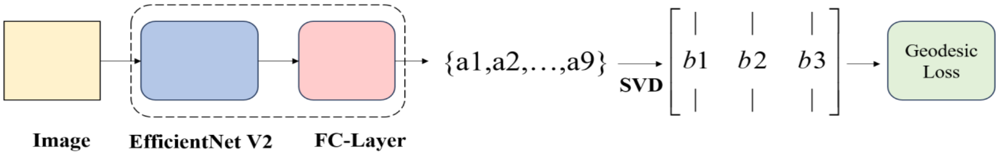

Figure 1. Overview of the proposed method

II. METHOD

A. Feature Extraction Network

Many existing neural networks use depthwise separable convolution to extract features, and although its structure possesses fewer parameters as well as smaller FLOPs compared to normal convolution, it is usually not able to fully utilize the gas pedal with the available hardware. This paper utilizes EfficientNet V2 as the feature extraction network, which is a more advanced and lightweight convolutional neural network model compared to EfficientNet. It is characterized by low number of parameters, high accuracy, and excellent training and inference speed. A notable improvement is the substitution of EfficientNet's shallow MBConv with the Fused-MBConv module. The Fused-MBConv module substitutes the expansion 1x1 convolution and depthwise 3x3 convolution in the primary branch of the original MBConv structure with a 3x3 convolution. This solves the problem of employing depth-separable convolution in the initial layer of the network. The issue of slowdown caused by using of depth-separable convolutions in the shallow layers of the network is effectively solved, resulting in an important enhancement in training speed. Figures 2 and 3 show the structure of the MBConv and Fused-MBConv modules, respectively. Table 1 shows the EfficientNet V2-S structure. The method proposed in this paper is adapted by changing the final fully connected layer output to 9.

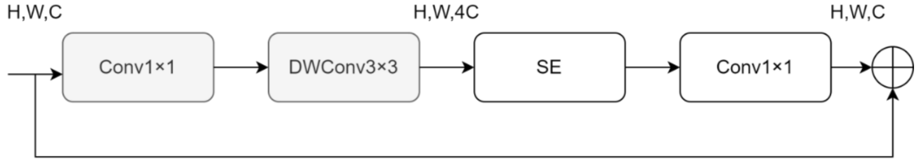

Figure 2. MBConv

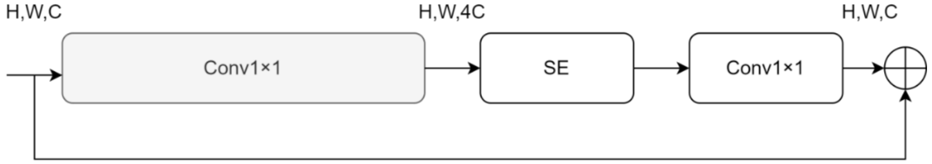

Figure 3. Fused-MBConv

TABLE I. EFFICIENTNETV2-S ARCHITECTURE

Stage	Operation	Stride	#Channels	#Layers
0	Conv3x3	2	24	1
1	Fused-MBConv1,3x3	1	24	2
2	Fused-MBConv4,3x3	2	48	4
3	Fused-MBConv4,3x3	2	64	4
4	MBConv4,3x3,SE0.25	2	128	6
5	MBConv6,3x3,SE0.25	1	160	9
6	MBConv6,3x3,SE0.25	2	256	15
7	Conv1x1&Pooling&FC	-	1280	1

B. R9+SVD

Choosing a suitable approach for representing rotation is vital for accurately estimating head posture. Traditionally, Euler angles have been employed. Nevertheless, this method of representing rotation is not ideal because to its susceptibility to gimbal lock. In such cases, specific sequences and angles of rotation can cause the loss of one of the three independent rotation axes. Another rotation representation is the quaternion method, which is not affected by gimbal lock but has the issue of double coverage. This means that for each rotation, there are two corresponding quaternions. While these two representations are physically equivalent, they exhibit a significant numerical discontinuity. Therefore, neural networks struggle to learn accurate poses in the presence of numerical discontinuities.

Figure 4 shows two examples of pictures with comparable visual presentation from the 300W-LP dataset. A comparison of the two pictures shows that the Euler angles and quaternions have distinct labeling values, notably the second value, yaw, in the Euler angles. Positive and negative numbers suggest entirely opposing attitudes. A better rotation representation is the rotation matrix, which is a continuous representation, only the rotation matrix can reflect the similarity of pose appearance. In $SO(3)$ space, the rotation matrix is a 3×3 matrix that satisfies the orthogonality criteria $RR^T = I$, where R^T represents the transpose of R , and I represents the identity matrix. The R9+SVD method can work with any 3×3 matrix and turn it into a valid rotation matrix in $SO(3)$. R9 refers to using the neural network to directly output 9 vector values that describe a 3×3 rotation matrix. The SVD method was created because directly estimating 9 variables might not work because rotation matrices have certain qualities, such as being orthogonal and having a positive determinant.

Euler Angles
$[-86.18 \quad 3.96 \quad -10.53]$
Quaternions
$[0.477 \quad -0.587 \quad 0.037 \quad -0.653]$
Rotation Matrix
$$\begin{bmatrix} 0.307 & 0.513 & 0.802 \\ -0.610 & -0.541 & 0.580 \\ 0.731 & -0.670 & 0.146 \end{bmatrix}$$

Euler Angles
$[-87.65 \quad -27.68 \quad 37.08]$
Quaternions
$[0.320 \quad 0.054 \quad -0.913 \quad -0.246 \ ]$
Rotation Matrix
$$\begin{bmatrix} -0.675 & 0.415 & 0.060 \\ 0.483 & 0.873 & 0.060 \\ 0.560 & -0.260 & -0.790 \end{bmatrix}$$

Figure 4. Image samples from 300W-LP dataset with different rotation representations

Given a 3×3 matrix, its singular value decomposition (SVD) is expressed as:

$$M = U\Sigma V^T \quad (1)$$

Here, U and V are 3×3 orthogonal matrices, and Σ is a diagonal matrix containing the singular values of matrix M .To project M onto the rotation matrix R , R must satisfy two conditions:

1) The column vectors of R must be of unit length and orthogonal to each other.

2) The determinant of R must equal 1.

Therefore, after adjusting the singular values, R is constructed as:

$$\Sigma^+ = diag(1,1,\det(UV^T)) \quad (2)$$

Reconstruct the rotation matrix using the adjusted singular value matrix:

$$R = U\Sigma^+ V^T \quad (3)$$

Hereby, $\det(UV^T)$ ensures that the determinant of R is 1, while the combination of U and V^T ensures that the column vectors of R are orthogonal. Therefore, the neural network predicts 9 parameters, which are then transformed into a 3 × 3 rotation matrix while sticking to the orthogonality requirement.

The advantages of this method are:

1) Smoothness: It provides a continuous and smooth representation, allowing optimization algorithms like gradient descent to converge effectively while avoiding issues such as the singularities of Euler angles or the double coverage problem of quaternions.

2) Robustness: SVD can be seen as a model architecture where the three column vectors of the matrix contribute equally to the prediction. This enhances robustness to input noise.

C. Geodesic Loss

The loss function commonly used in previous head pose estimation tasks is the L2 loss function, and the calculation method is equation (4). However, in the head pose estimation task, there are some problems when using the L2 loss function to measure the difference between rotations. First, the L2 loss does not take into account the periodicity of the rotation angle, which makes it impossible to correctly evaluate the similarity between rotations close to 360 degrees or −360 degrees. Secondly, the L2 loss assumes that all dimensional changes are independent and linear, which is inconsistent with the geometric structure of the rotation matrix or quaternion.

$$loss(x, y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \quad (4)$$

The geodesic loss function measures the distance between two rotation matrices along the shortest path on the manifold, known as the geodesic. The geodesic loss function is calculated based on the trace of the rotation matrices, with the formula given as:

$$d(R_1, R_2) = \cos^{-1}(\frac{tr(R_1 R_2{}^T) - 1}{2}) \quad (5)$$

R_1 and $R_2 \in SO(3)$, representing the predicted rotation matrix and the true rotation matrix, respectively. The trace (tr) denotes the sum of the diagonal elements of a matrix. This distance will be used as the loss function for the neural network in subsequent experiments.

III. EXPERIMENTS AND RESULTS

A. Datasets

This work trained and evaluated it method on various types of datasets. The most commonly used publicly available datasets for head pose estimation are 300W-LP [15], AFLW2000[16], and BIWI [17].

1) The 300W-LP dataset consists 66,225 facial pictures, which are increased to 122,450 samples using image flipping augmentation. It encompasses a diverse array of postures and comprehensive 3D annotation data. The ground truth is given as Euler angles, which were transformed into matrix representation following the method described by Hempel [18].

2) The AFLW2000 dataset includes the

initial 2,000 face photos that were chosen from the AFLW dataset. These images are accompanied by 68 key point annotations. It includes a range of face positions, including various degrees of rotation and emotions.

3) The BIWI dataset includes video sequences of 24 individuals, totaling 15,678 images. Each frame provides detailed 3D head pose and key point annotations, covering various head pose variations in real-world scenarios. The MTCNN [19] facial detection algorithm was used to extract the head region from the images.

B. Evaluation Metrics

The head pose estimate error is measured using the Mean Absolute Error (MAE) of Euler angles, which is the most widely used metric. This is represented by Equation (6).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(|\, x_g - x_p \,|) \qquad (6)$$

N refers to the total number of face images, x_g represents the true values of the head poses, and x_p represents the predicted values of the head poses.

C. Implementation Details and Results

This work employed PyTorch to create the whole model, with EfficientNetV2-S serving as the backbone network, and trained the network for 30 epochs with the Adam optimizer. The initial learning rates for the backbone network and the final fully connected layer were set to 1e-5 and 1e-4, respectively, with each learning rate halving every 10 epochs. The batch size was set at 64.

In the first experiment, the network was trained using the synthetic 300W-LP dataset and subsequently tested on two real-world datasets: AFLW2000 and BIWI. The evaluation metric used was the mean absolute error (MAE) of Euler angles, which required transforming the predicted rotation matrices into Euler angles for comparison purposes. Table 2 presents the results of the first experiment, comparing the proposed approach to other state-of-the-art landmark-free head pose estimation methods. The experimental results demonstrate that the proposed strategy outperformed the current best methods by approximately 22% and achieved the lowest error rates in pitch, yaw, and roll angles on the AFLW2000 dataset. On the BIWI dataset, the approach exceeded seven out of eight of the most advanced algorithms in terms of MAE. Figure 5 illustrates the results of the method on the AFLW2000 dataset after converting the predicted rotation matrices to Euler angles.

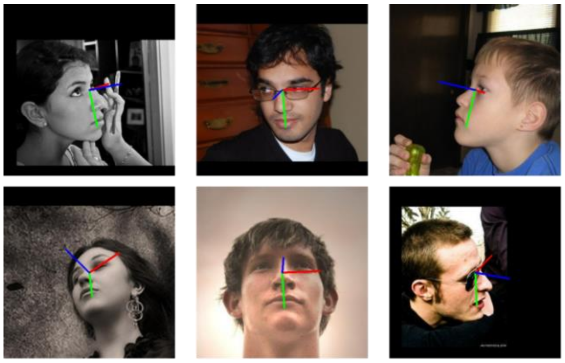

Figure 5. Example images of Euler angle visualization using rotation matrix transformation from AFLW2000 dataset

TABLE II. COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE AFLW2000 AND BIWI DATASET

	AFLW2000				BIWI			
Models	**Yaw**	**Pitch**	**Roll**	**MAE**	**Yaw**	**Pitch**	**Roll**	**MAE**
HopeNet[4]	6.40	6.53	5.39	6.11	4.54	5.15	3.37	4.36
FSA-Net[8]	4.50	6.08	4.64	5.07	4.64	5.61	3.57	4.61
HPE[6]	4.80	6.18	4.87	5.28	3.12	5.18	4.57	4.29
QuatNet[5]	3.97	5.62	3.92	4.50	2.94	5.49	4.01	4.15
WHENet[7]	5.11	6.24	4.92	5.42	3.99	4.39	3.06	3.81
TriNet[9]	4.04	5.77	4.20	4.67	4.11	4.76	3.05	3.97
FDN[10]	3.78	5.61	3.88	4.42	4.52	4.70	2.56	3.93
6DRepNet[18]	3.63	4.91	3.37	3.97	3.24	4.48	2.68	3.47
9D-EfficientNet	3.57	4.69	3.28	3.85	4.08	4.17	2.94	3.73

In the second experiment, the method outlined by FSA-Net was followed, with the BIWI dataset randomly split into training and testing sets in a 7:3 ratio. The results were compared with other networks that employed the same experimental approach. Table 3 presents the results of the second experiment. The proposed method outperforms other methods in terms of MAE, and shows superior performance in yaw and pitch, with roll being better than most. These experimental results demonstrate the robustness of the proposed method, as it achieves stable and accurate results in both Euler angle and MAE across different datasets.

TABLE III. EULER ERROR COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE 70/30 BIWI DATASET

	BIWI			
Models	**Yaw**	**Pitch**	**Roll**	**MAE**
HopeNet[4]	3.29	3.39	3.00	3.23
FSA-Net[8]	2.89	4.29	3.60	3.60
TriNet[9]	2.93	3.04	2.44	2.80
FDN[10]	3.00	3.98	2.88	3.29
MDFNet[20]	2.99	3.68	2.99	3.22
DDD-Pose[21]	3.04	2.94	2.43	2.80
6DRepNet[18]	2.69	2.92	2.36	2.66
9D-EfficientNet	2.62	2.36	2.51	2.50

To demonstrate the superiority of the geodesic loss function as a distance metric for head pose estimation, additional tests were conducted using the rotation matrix. To support this claim, the previous experiments were replicated by training the network with the L2 loss function. Table 4 presents the effectiveness of the proposed technique when trained with two distinct loss functions. Training the network with the geodesic loss function yields superior results compared to the L2 loss.

TABLE IV. COMPARISON OF THE MAE BETWEEN L2 AND GEODESIC LOSS

	AFLW2000	**BIWI**	**70/30 BIWI**
Loss function	MAE	MAE	MAE
L2 Loss	3.90	3.92	2.71
Geodesic Loss	3.85	3.73	2.50

This paper also examines the influence of different backbone networks on the results obtained by employing geodesic loss. In order to do a comparison, this research employed the ResNet [21] network as an illustrative example. The findings shown in Table 5 demonstrate that the approach outlined in this research achieves exceptional results when used to the EfficientNet V2-S backbone network. By employing ResNet18 as the foundation network, this approach surpasses the majority of previous approaches in terms of performance on both the AFLW2000 and BIWI datasets. This illustrates that employing a suitable rotation representation greatly enhances the accuracy of head pose estimation.

TABLE V. COMPARISON OF MAE BETWEEN RESNET AND EFFICIENTNETV2 BACKBONE NETWORKS

	AFLW2000	**BIWI**	**70/30 BIWI**
Models	**MAE**	**MAE**	**MAE**
ResNet18	4.37	3.70	2.64
EfficientNetV2-S	3.85	3.73	2.50

In the final experiment, the THOP library was used to compare the proposed method with 6DRepNet in terms of parameter count and floating-point operations (FLOPs). As shown in Table 6, the proposed approach achieved a lower MAE while requiring fewer parameters and FLOPs.

TABLE VI. COMPARISON OF PARAMETERS AND FLOPS BETWEEN 6DREPNET AND OUR METHOD

Models	**Params**	**FLOPs**
6DRepNet	43.752M	9.844G
9D-EfficientNet	20.189M	2.901G

IV. CONCLUSIONS

In this research, provide an appearance-based, unconstrained, end-to-end head posture estimation approach. Following the assumption that rotation matrices are better suited to deep learning in 3D rotation problems, and provide a continuous 9D vector + SVD technique for head pose estimation. In addition, this research uses the geodesic loss function rather than the usual MSE to better correspond with the rotation matrix representation. Experiments show that using the EfficientNetV2 backbone network, this approach surpasses other most advanced methods on the AFLW2000 dataset and most methods on the BIWI dataset. In further experiments, this investigated the effects of various loss functions and backbone networks on the findings, as well as comparisons of parameter count and floating-point operations. All of the experiments show that this approach is robust, reliable, and lightweight.

REFERENCES

[1] Strazdas Dominykas, Hintz Jan, AlHamadi Ayoub. Robo-hud: interaction concept for contactless operation of industrial cobotic systems [J]. Applied Sciences, 2021, 11(12):5366-5366.

[2] CHARISSIS V, FALAH J, LAGOO R, et al. Employing emerging technologies to develop and evaluate in-vehicle intelligent systems for driver support: infotainment ar hud case study [J]. Applied Sciences, 2021,11(4):1397-1397.

[3] Werner, P., Saxen, F., & Al-Hamadi, A. Landmark based head pose estimation benchmark and method. In ICIP, 2017.

[4] Ruiz, N., Chong, E., & Rehg, J. M. Fine-grained head pose estimation without keypoints. In CVPR, 2018.

[5] Hsu H W, Wu T Y, Wan S, et al. QuatNet: Quaternion-Based Head Pose Estimation with Multiregression Loss [J]. Multimedia, IEEE Transactions on, 2018. DOI:10.1109/TMM.2018.2866770.

[6] Huang B, Chen R, Xu W, et al. Improving head pose estimation using two-stage ensembles with top-k regression [J]. Image and Vision Computing, 2019, 93.DOI:10.1016/j.imavis.2019.11.005.

[7] Zhou Y, Gregson J. WHENet: Real-time Fine-Grained Estimation for Wide Range Head Pose [J]. 2020. DOI:10.48550/arXiv.2005.10353.

[8] Yang T Y, Chen Y T, Lin Y Y, et al. FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation from a Single Image [J]. IEEE, 2020. DOI:10.1109/CVPR.2019.00118.

[9] Cao Z, Chu Z, Liu D, et al. A Vector-based Representation to Enhance Head Pose Estimation[C]//Workshop on Applications of Computer Vision. IEEE, 2021. DOI:10.1109/WACV48630.2021.00123.

[10] Zhang H, Wang M, Liu Y, et al. FDN: Feature Decoupling Network for Head Pose Estimation [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7):12789-12796. DOI:10.1609/aaai.v34i07.6974.

[11] Dhingra N. LwPosr: Lightweight Efficient Fine-Grained Head Pose Estimation [J]. arXiv e-prints, 2022. DOI:10.48550/arXiv.2202.03544.

[12] Zhou Y, Barnes C, Lu J, et al. On the Continuity of Rotation Representations in Neural Networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019. DOI:10.1109/CVPR.2019.00589.

[13] Geist A R, Frey J, Zobro M, et al. Learning with 3D rotations, a hitchhiker's guide to SO (3) [J]. arxiv preprint arxiv:2404.11735, 2024.

[14] Tan M, Le Q V. EfficientNetV2: Smaller Models and Faster Training [J]. 2021. DOI:10.48550/arXiv.2104.00298.

[15] Xiangyu Zhu, Zhen Lei, Xiaoming Liu 0002, et al. Face alignment across large poses: a 3d solution. [J]. CoRR, 2015.

[16] Zhu X, Lei Z, Yan J, et al. High-fidelity Pose and Expression Normalization for face recognition in the wild [J]. IEEE, 2015. DOI:10.1109/CVPR.2015.7298679.

[17] Fanelli G, Matthias Dantone. Random Forests for Real Time 3D Face Analysis [J]. International Journal of Computer Vision, 2013, 101(3):437-458. DOI:10.1007/s11263-012-0549-0.

[18] Hempel T, Abdelrahman A A, Al-Hamadi A .6D Rotation Representation for Unconstrained Head Pose Estimation [J]. arXiv e-prints, 2022. DOI:10.48550/arXiv.2202.12555.

[19] Zhang K, Zhang Z, Li Z, et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks [J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503. DOI:10.1109/LSP.2016.2603342.

[20] Liu H, Fang S, Zhang Z, et al. MFDNet: Collaborative Poses Perception and Matrix Fisher Distribution for Head Pose Estimation [J]. IEEE Transactions on Multimedia, 2021, PP (99): 1-1. DOI:10.1109/TMM.2021.3081873.

[21] Aghli N, Ribeiro E. A Data-Driven Approach to Improve 3D Head-Pose Estimation[C]//International Symposium on Visual Computing. Springer, Cham, 2021. DOI:10.1007/978-3-030-90439-5_43.

[22] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition [J]. IEEE, 2016. DOI:10.1109/CVPR.2016.90.