Journal of King Saud University - Computer and Information Sciences 36 (2024) 101869

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University - Computer and Information Sciences

journal homepage: www.sciencedirect.com

Full length article

DADL: Double Asymmetric Distribution Learning for head pose estimation in wisdom museum

Wanli Zhao [a,1], Shutong Wang [a,b,1], Xiaoguang Wang [a,e,*], Duantengchuan Li [c,*], Jing Wang [d,e], Chenghang Lai [f], Xiaoxue Li [g]

[a] School of Information Management, Wuhan University, Wuhan 430072, China
[b] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China
[c] School of Computer Science, Wuhan University, Wuhan 430072, China
[d] School of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[e] Intellectual Computing Laboratory for Cultural Heritage, Wuhan University, Wuhan 430072, China
[f] School of Computer Science, Fudan University, Shanghai 200438, China
[g] Yichang Education Information Technology Center, Yichang 443000, China

ARTICLE INFO

Keywords:
Head pose estimation
Label distribution learning
Gaussian distribution
Asymmetric
Feature similarity

ABSTRACT

Head pose estimation plays a pivotal role in various applications, including augmented reality and human–computer interaction within intelligent museum environments. Head pose estimation conventionally relies on hard labels. However, acquiring the "ground truth" through subjective means introduces an element of uncertainty into the labels for head pose estimation. The introduction of soft labels offers a potential remedy for this uncertainty. However, existing head pose estimation methods based on soft labels neglect the asymmetry of head pose. After careful observation, two types of asymmetry have been identified in human head pose: within angle and between angle asymmetry. Taking these two characteristics into account, we have devised a Double Asymmetric Distribution Learning (DADL) network model for the precise estimation of head pose angles. This model employs distinct soft label distribution mechanisms to capture within-angle and between-angle nuances in head pose variations. Thereby enhancing the interpretability, generalization capability, and classification accuracy of head pose estimation models. Extensive experiments were conducted on various widely recognized benchmarks, including the AFLW2000 and BIWI datasets. The results substantiate substantial advantages of our model over conventional approaches.

1. Introduction

Head pose estimation (HPE) has obtained increasing attention in recent years, as it plays an important role in many fields, such as behavior analysis (Kumar et al., 2017), gaze estimation (Ranjan et al., 2017), virtual reality (Xu et al., 2018), intelligence education (Chen et al., 2014; Song et al., 2023) and human–computer interaction in museums (Banfi et al., 2023). It assumes that the human head is a rigid object and contains three degrees of freedom (DOF) in the head pose, namely yaw, pitch and roll. Existing methods for HPE tasks can be roughly divided into two categories: (1) hard label based HPE, and (2) soft label based HPE.

The hard label methods have higher requirements for the accuracy of the ground-truth pose. However, in previous studies, the "ground truth" pose is often obtained in a more subjective way. For example, the Pointing'04 head pose database is collected by requiring human subjects to sit in the same position in a room and stare at 93 markers attached to different positions of the room (Gourier et al., 2004). This method can only collect rough poses due to the following two points: First, the subject's head is not guaranteed to be in exactly the same

* Corresponding authors.
E-mail addresses: wanlizhao146@gmail.com (W. Zhao), wxguang@whu.edu.cn (X. Wang), dtclee1222@gmail.com (D. Li).
[1] These authors contributed equally to this work.
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2023.101869
Received 6 August 2023; Received in revised form 1 November 2023; Accepted 4 December 2023
Available online 14 December 2023

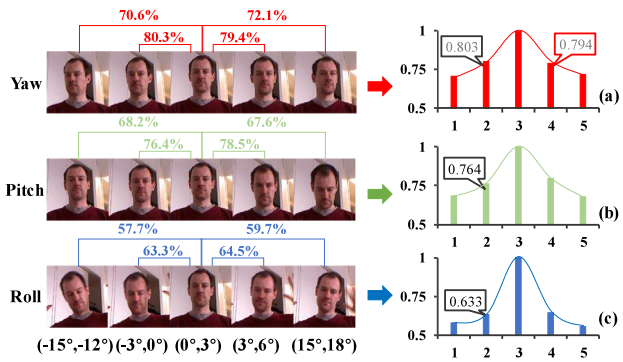

Fig. 1. Characteristics of asymmetry of the head pose estimation task. (a), (b), (c) indicate the similarity change curves of the rotated image and the center image on Yaw, Pitch, and Roll angles, respectively.

position in the 3D space. Second, people cannot point their head at the marker very accurately. The Pandora head pose database is labeled by a wearable Inertial Measurement Unit (IMU) sensor, which has been worn in an invisible position by the human subjects (Borghi et al., 2017). Similarly, owing to limitations of the sensor (e.g., equipment deviations), the noise in the pose label (i.e., label ambiguity) still exists. Regarding this issue, Geng et al. think that it is better to use soft labels to represent the pose of face images (Geng et al., 2022). Different from explicit hard labels, soft labels make the description degrees of all poses form a data form similar to a multivariate probability distribution. Therefore, such soft label of head poses is also known as label distribution learning (LDL). Compared with hard labels, LDL provides a more flexible and powerful label representation. Meanwhile, it also better represents the complexity and uncertainty of the sample to solve the problem of label ambiguity. The results of several studies have shown that the LDL-based methods show excellent performance in HPE tasks.

However, there are two important properties that have not been considered in the HPE task based on the LDL method. The first property, which we call within angle asymmetry (WAA), is that the similarity of the head pose images varies across the three angles of yaw, pitch and roll for the same pose angle rotated on a fixed central pose. As shown in Fig. 1, taking the yaw dimension as an example, when the front face image is turned to the left and to the right by the same angle respectively, their two similarities are different, with values of 0.803 and 0.794 respectively. The second property is called between angle asymmetry (BAA), in the same angle, under the same deflection, the feature similarity of a single dimension will also change. As shown in Fig. 1, taking the yaw, pitch and roll dimension as an example, the changes of similarity are different when the front face images are turned to the left by the same angle consecutively, with values of 0.803, 0.764 and 0.633 respectively.

Considering the angle asymmetry properties in the LDL method, we propose a novel double asymmetric distribution learning for head pose estimation (DADL) based on the discovered head pose asymmetry phenomenon. After careful similarity analysis, we found that similarity difference is very small in WAA. Therefore, we decide to use three different standard Gaussian distributions to fit the WAA property to avoid the introduction of too many model parameters to reduce the generalization ability of the model, which can also improve the training efficiency of the model. The main contributions of this work compared to previous studies can be summarized in three aspects:

- A DADL head pose estimation framework based on Gaussian-based distribution learning is proposed. The model is trained with RGB face images. To the best of our knowledge, this is the first work for HPE with double asymmetric multi-angle distribution learning.
- In label distribution construction, we use three different Gaussian distributions to construct soft labels to fit the model, considering the dual asymmetry of the human head pose between and within angles, and for the label blurring and model overfitting problems caused by hard labels.
- Our model is trained on the sizeable synthetic dataset 300W-LP, tested on multiple datasets AFLW2000 and BIWI. The proposed method achieves the preferable performance on both prediction accuracy and robustness.

This paper is structured as follows. Related work is presented in Section 2. The proposed DADL model for HPE and its corresponding module, including soft-label distribution and hard label distribution, are described in Section 3. The experiment results, comparison, analysis and ablation study are elaborately stated in Section 4. Finally, Section 5 summarizes the presented work and prospects of future work.

2. Related work

2.1. Head pose estimation

Convolutional neural networks are widely used in various fields (Wang et al., 2023b; Li et al., 2023b; Wang et al., 2023a), especially for image feature extraction. Head pose estimation is the process of analyzing human head orientation from RBG images (Li et al., 2023a), which can be automatically trained in an end-to-end manner on large-scale datasets by using convolutional neural networks (CNNs) (Abdullah et al., 2013; Murtza et al., 2017, 2019). It can be divided into three angles according to their directions of head motion: pitch (forward to backward movement of the neck), roll (right to left rotation of the head), and yaw (right to left bending of the neck) (Thai et al., 2022). Head pose estimation can be achieved in two ways. One way is to use a fixed label value to predict the head pose (i.e., Hard Label Learning). The other way is to utilize the probability distribution of the label to predict the head pose (i.e., Label Distribution Learning).

2.2. Hard label learning based HPE approaches

Among these hard label learning based HPE approaches, early work commonly used landmark-based methods to predict head pose. These methods mainly estimate the head pose by calculating the correspondence between 2D facial landmarks and 3D head models. For example, (Sun et al., 2013) adopted a three-layer convolutional network to estimate the position of facial landmarks, whereas this method does not achieve the desired results in every task. Therefore, (Kumar et al., 2017) developed the keypoint estimation and pose prediction of unconstrained faces by learning efficient H-CNN regressors (KEPLER) method that can obtain precise landmarks localization on unconstrained faces. However, this method cannot meet the requirements of stability, accuracy and real-time performance in complex scenes. To solve the problem, (Ranjan et al., 2017) proposed a multi-task deep learning method called HyperFace to simultaneously detect faces, locate landmarks and estimate head pose.

Nevertheless, the performance of landmark-based methods is highly dependent on the accuracy of facial landmark detection, which is not applicable to conditions where landmarks are invisible. To tackle this problem, increasing works use landmark-free methods for HPE prediction task. These methods directly estimate the head pose from images without facial landmarks. For example, (Ruiz et al., 2018) exploited a multi-loss deep network to directly predict intrinsic Euler angles (yaw, pitch and roll) from image intensities. (Zhou and Gregson, 2020) proposed a wide head pose estimation network (WHENet) method that can capture more comprehensive head movement information and enable real-time and fine-grained estimation of head pose. (Cao et al., 2021) proposes a vector-based head pose estimation method, which can solve the discontinuity issues caused by Euler angles. However, the

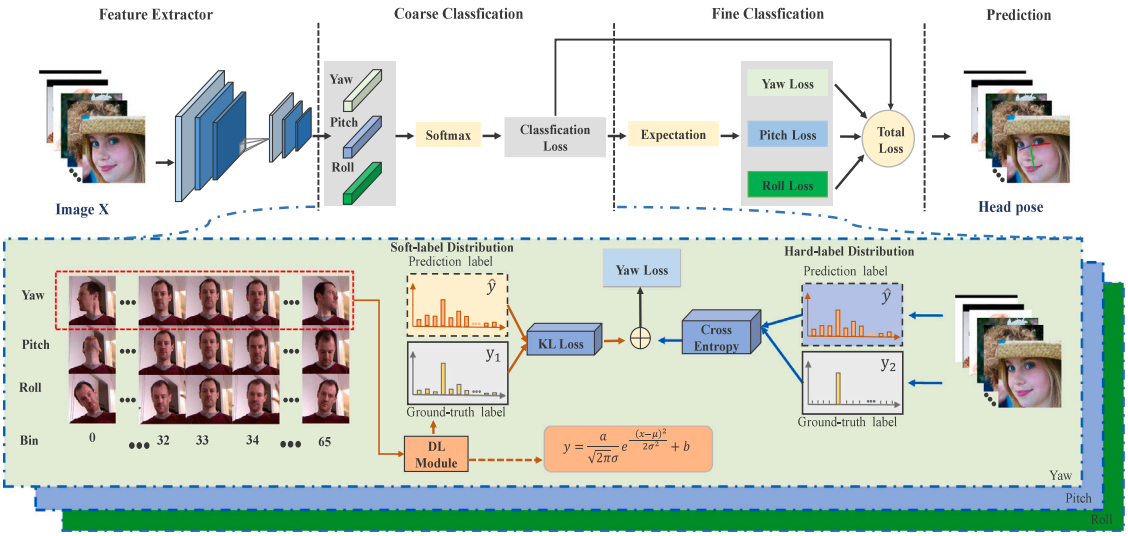

Fig. 2. The architecture of the proposed DADL model. It mainly consists of two stages: coarse and fine classification. In the first stage we performed soft label distribution construction and soft label distribution learning, combined with the corresponding hard label distribution, and performed coarse-grained category regression in that stage. The second stage uses finer-grained losses based on coarse categorization to regress head pose angles respectively.

methods mentioned above (i.e., landmark-free methods) have several limitations. For instance, landmark-free methods often have higher computational overhead and require a mass of data to drive network learning. Moreover, landmark-free methods do not consider the probability problem of non-category labels, which may lead to the problem of label ambiguity.

2.3. Label distribution learning based HPE approaches

Given the limitation of hard label learning based HPE approaches, several researchers use label distribution learning (LDL) methods to predict head pose. The LDL method describes each label as a probability distribution rather than a binary classification, that is, a label distribution covers multiple class labels, and denotes the relative importance to which each label describes the instance (Jia et al., 2019). The basic idea of LDL is to provide partial weights for neighboring labels to introduce strong correlations of observations, meanwhile, neighboring labels provide rich cues for extracting discriminative features (Geng, 2016). Compared with the hard label learning based HPE approaches, The LDL method have two advantages. First, it is a novel machine learning framework that can be used to solve the problem of label noise and ambiguity (Xu et al., 2021). Second, it effectively relaxes the requirement for a large number of training images, that is, the training examples for each head pose can be reinforced without actually increasing the total training examples (Geng et al., 2022).

Due to the above advantages, the label distribution learning technique is widely used in image classification and regression tasks, such as facial age estimation (Wen et al., 2020), facial expression recognition (Zhao et al., 2021), and head pose estimation (Liu et al., 2021). By combining the two concurrent processes label distribution refinement and slack regression refinement, (Li et al., 2020) employed label refinery network (LRN) to estimate facial age from color images. Based on intensity label distribution, (Chen et al., 2022) presented a novel facial expression analysis method to analyze the children's empathy ability, and they used both the categories and the intensities of facial expressions to detect children's emotional states. To solve the problem of inaccurate pose labels, (Geng et al., 2022) proposed a hierarchical multivariate label distribution (HMLD) method to deal with fine-grained head pose estimation.

3. Proposed DADL model

In this section, we first describe the overall framework of our model, then elaborate on each part of the model, and finally describe the optimization process of the model.

3.1. Outline of DADL

Our proposed DADL network architecture is shown in Fig. 2, which consists of the following four fundamental modules: feature extractor module, coarse classification module, fine classification module as well as prediction model, where the coarse classification stage principally includes the soft-label distribution construction and the hard-label distribution learning module. The idea behind this design is that by performing bin rough classification, we use a very stable softmax layer, cross entropy, and Kullback–Leibler (KL) divergence loss, so network learning predicts the neighborhood of the pose in a robust way. Through a composite loss with three cross entropy and KL, one signal per Euler angle, three signals are back-propagated into the network, thereby improving learning. In order to obtain fine-grained predictions, we calculate the expectation of each output angle of bin.

3.2. Feature extractor

In the task of head pose estimation, a feature extractor can extract useful features such as edges, shapes, and textures from the raw pixel data. Common feature extractors include CNN (LeCun et al., 1989), RNN (Zaremba et al., 2014), and Autoencoder (Rumelhart et al., 1986). These algorithms learn and optimize parameters to extract the most representative or discriminative features, thereby achieving accurate extraction of head pose. Our model adopts the widely used backbone model ResNet50 (He et al., 2016a) for extracting head pose information.

3.3. Coarse classification

To provide a detailed explanation of this module, we first conducted a rough classification of the head pose range. Then, we elaborated on the BAA characteristics in detail. Finally, based on these BAA characteristics, we constructed a soft label distribution while also preserving the construction part of the hard label.

Fig. 3. Samples classification of head pose anisotropy. The first, second, and third rows represent samples from yaw, pitch, and roll, respectively. The last row, Bin, indicates the range of the category to which the current image belongs. We divide the samples within plus or minus 99° according to 3°, and the divided category is denoted as Bin. For instance, Bin equals to 33 means that the current image is a frontal face image.

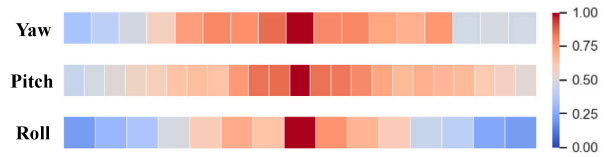

Fig. 4. Asymmetric similarity distribution matrix. We compute the all similarity between center category pose and other category poses in BIWI dataset.

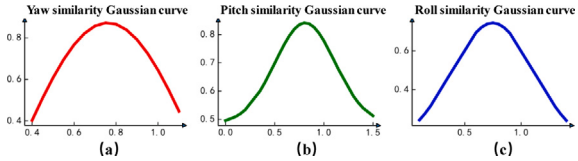

Fig. 5. Asymmetric similarity distribution curve. The shape of similarity distribution curve is fitted by a one-dimensional Gaussian distribution. (a), (b), (c) denote the similarity fitting curves for yaw, pitch and roll respectively.

3.3.1. Bins partition details

According to our statistics, the range of different angles variation in the head pose datasets does not exceed −99° and +99°. Therefore, we divide all the head pose angle categories (pitch, yaw, roll) into the following value ranges: [−99°, −96°, −93°, …, 0°, 3°, …, +96°, +99°], so that the angle difference between adjacent head pose categories is 3°, then the granularity $\tau = 3$. The angular categories bins after division are: [0, 1, 2, …, 33, 34, …, 64, 65].

3.3.2. Between angle asymmetry (BAA)

Our human head is roughly considered a symmetric non-spherical rigid body, which implies that the variation of head rotation varies in different directions. In other words, for a certain head pose image X, it is rotated by a fixed angle in the directions corresponding to pitch angle, yaw angle, and roll angle, respectively, and the rotated image is denoted as X_{yaw}, X_{pitch}, X_{roll}. However, the similarity between the original image X and the rotated images X_{yaw}, X_{pitch}, and X_{roll} are different. We named this phenomenon as the between angle asymmetry (BAA). The details are as flows:

As shown in Fig. 3, the original image with head pose X_o(yaw=33, pitch=33, roll=33) as an example, it can be observed that the face shape difference between X_o and X_y (yaw=0, pitch=33, roll=33), the face shape difference between X_o and X_p (yaw=33, pitch=0, roll=33) face shape differences and face shape differences between X_o and X_r (yaw=33, pitch=33, roll=0) are different. To quantitatively formulate this phenomenon, the cosine similarity, which is generally used in natural language processing, is introduced to calculate the feature similarity of two images. LeNet is effective in computer vision to extract the texture and edge features of the underlying image, so a pre-trained LeNet neural network is used to extract the image features. The vector generated by the last full-connection (FC) layer is the most representative feature of the figure. Thus, the vector of the last full-connection layer from the pre-trained neural network is obtained to calculate the feature similarity of two vectors to compare the feature similarity of two images. The whole process of calculating the image features is denoted as $F(\cdot)$. Given two head pose images X_1 and X_2, the head pose feature similarity is defined as follows,

$$S(X_1, X_2) = Similarity(F(X_1), F(X_2)) = \frac{F(X_1) \cdot F(X_2)}{\| F(X_1) \| \times \| F(X_2) \|}, \quad (1)$$

where $F(X)$ denotes the feature vector of image X. Eq. (1) is used to compute the similarity between the two pose images, which in turn validates the presence of asymmetry in the head pose. It is not difficult to find that the head pose angles are of natural continuity, and faces close to the target pose look very similar. Therefore, additional knowledge about faces in different poses can be introduced to enhance the learning problem. To put it simply, it is to learn a specific pose while using the information of faces in adjacent poses. To achieve this, we calculate the feature similarity of any category of head pose images with all the remaining categories of images, and several feature similarity matrices can be obtained. As shown in Fig. 3, we take the front face image X (yaw=0°, pitch=0°, roll=0°) as an example, calculating the similarity between this image and the rest of the head pose bin images by the formula $S(X_1, X_2)$, we can get the following matrix in Fig. 4.

We can fit the matrix using three different Gaussian distributions, and the resulting curves are shown in Fig. 5(a), (b) and (c). Here, (a), (b), and (c) represent the Gaussian distribution curves fitted by the feature similarity values for the yaw, pitch, and roll angles, respectively. We can observe that the three curves are generally bell-shaped and symmetrical and have their highest probability values at the midpoint. The similarity distribution curve for pitch is flatter, indicating that the similarity changes for this angle are smoother. Additionally, we can observe that the curves fitted by the similarity distribution show an asymmetric characteristic.

We treat each head pose image label as a Gaussian distribution rather than a traditional hard label, where adjacent head pose images can provide supplementary information for the target image. Specifically, since each label of the head pose image based on the Gaussian distribution describes the similarity between the current image and the adjacent category head pose images, the Gaussian distribution label covers the feature information of both the current and adjacent category head pose images. This may not only help to learn the true pose of the facial image but also help to learn its adjacent poses. To strengthen the entire learning process, we use three different Gaussian label distributions to describe the yaw angle, pitch angle, and roll angle in the face sample.

3.3.3. Within angle asymmetry (WAA)

The similarity of the head pose images varies across the three angles of yaw, pitch and roll for the same pose angle rotated on a fixed central pose. However, when the front face image is turned to the left and to the right by the same angle respectively, their two similarities are different. We named this phenomenon as the Within Angle Asymmetry (WAA). However, different with BAA, the difference of similarities is very small. Taking into account that the small differences have minimal impact on our model, we decide to use three different standard Gaussian distributions to fit the WAA property. This will reduce the introduction of parameters, make the construction of the model simple and efficient, and improve the generalization ability of the model.

3.3.4. Soft-label distribution

Based on the category cosine similarity values of the head pose, the parameters of the head pose similarity distribution obeying the proposed Gaussian distribution can be analyzed, and the Gaussian distribution is applied to the construction of the head pose label distribution. Formally, the construction for the correlation emotion label distribution by a mapping function that maps the stored pose labels and values to the head pose label distribution D. The process is seen below: given a training set $E = (X_0, \mathbf{q}_0), (X_1, \mathbf{q}_1), \ldots, (X_n, \mathbf{q}_n)$, where $\mathbf{Q}_i = \mathbf{q}_0, \mathbf{q}_1, \ldots, \mathbf{q}_n$ is the ground-truth hard-label to X_i, n stands for the size of the batch size. The label of the reconstructed X_i can be expressed as

$$f_{X_i} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{2}$$

The labels under the yaw, pitch and roll dimensions can all be represented by the 3.3 formula. Where $\mu = 0.69$ and $\sigma = 0.17$ in the yaw dimension, $\mu = 0.68$ and $\sigma = 0.14$ in the pitch dimension, $\mu = 0.51$ and $\sigma = 0.23$ in the roll dimension. Here we take the yaw angle as an example to illustrate the soft tag distribution. Given a face image X_i a complete set of yaw angle labels $\mathbf{y} = y_1, y_2, \ldots, y_n$, if its yaw angle labels are y_i, $i = 1, 2, \ldots, n$, then the corresponding yaw angle labels are distributed as $\mathbf{D}^{y}_{X_i} = \{d^{y_1}_{X_i}, d^{y_2}_{X_i}, \ldots, d^{y_n}_{X_i}\}$, with the m-th dimension as follows:

$$d^{y_j}_{X_i} = \frac{Y^{y_j}_{X_i}}{\sum_{k=1}^{n} Y^{y_k}_{X_i}}, \tag{3}$$

$$Y^{y_j}_{X_i} = f^{yaw}_{X_i} = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right), \tag{4}$$

where j represents the yaw angle of the j-th bin classification, while y denotes the same head pose label as q_i, μ and σ represent the expectation and variance of the Gaussian function in the yaw angle. Thus $d^{y_j}_{X_i}$ denotes the extent to which the labels describe the example X_i under the constraint $\sum_{k=1}^{n} d^{y_k}_{X_i} = 1$, which means that the set of labels y completely describes the example. The other two angular label distributions follow the same definition but have different parameter constraints. $\mathbf{D}^{p}_{X_i} = \{d^{p_1}_{X_i}, d^{p_2}_{X_i}, \ldots, d^{p_n}_{X_i}\}$ and $\mathbf{D}^{r}_{X_i} = \{d^{r_1}_{X_i}, d^{r_2}_{X_i}, \ldots, d^{r_n}_{X_i}\}$ can be obtained by a set of pitch angle labels $\mathbf{p} = p_1, p_2, \ldots, p_n$ and roll angle labels $\mathbf{r} = r_1, r_2, \ldots, r_n$ of X_i, respectively.

$$d^{p_j}_{X_i} = \frac{Y^{y_j}_{X_i}}{\sum_{k=1}^{n} Y^{y_j}_{X_i}}, \tag{5}$$

$$Y^{p_j}_{X_i} = f^{pitch}_{X_i} = \frac{1}{\sqrt{2\pi}\sigma_p} \exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right), \tag{6}$$

$$d^{r_j}_{X_i} = \frac{Y^{y_j}_{X_i}}{\sum_{k=1}^{n} Y^{y_j}_{X_i}}, \tag{7}$$

$$Y^{r_j}_{X_i} = f^{roll}_{X_i} = \frac{1}{\sqrt{2\pi}\sigma_r} \exp\left(-\frac{(x-\mu_r)^2}{2\sigma_r^2}\right), \tag{8}$$

According to the distribution curves after similarity fitting, we set the parameters on the two datasets AFLW2000 and BIWI with variance $\sigma^A_y = 1$, $\sigma^A_p = 2$, $\sigma^A_r = 0.5$ and $\sigma^B_y = 0.35$, $\sigma^B_p = 0.6$, $\sigma^B_r = 0.3$, respectively, while the mean $\mu_y = \mu_p = \mu_r = 0$. Consequently, the training set can be represented as $\{(X_i, (\mathbf{D}^{y}_{X_i}, \mathbf{D}^{p}_{X_i}, \mathbf{D}^{r}_{X_i})), 1 \le i \ge n\}$. The goal of learning is to train a set of network parameters θ to generate probabilistic label distributions.

3.3.5. Hard-label distribution

Based on the earlier classification of bins, we know the number of predicted categories for the coarse classification module. Following the one-hot encoding strategy, each category label is encoded to form a hard label distribution. According to the Bins Partition Details, we know the angular categories bins after division are: [0, 1, 2, …, 33, 34, …, 64, 65]. Then the [+96°, +99°] bin's label distribution is [0, 0, 0, …, 0, 0, …, 0, 1].

3.4. Fine classification

By using bin classification, the coarse classification module can obtain an accurate category prediction. However, in the head pose task, the goal is to obtain accurate angle values rather than a category. The fine classification module calculates the expected value based on the coarse classification category to achieve the goal of fine classification. The specific calculation formula is as follows:

$$angle\_predicted = bin \cdot 3 - 99, \tag{9}$$

where bin stands for the predicted bin and $angle\_predicted$ the predicted value.

3.5. Prediction of DADL model

In this study, to encourage the model to learn both the correct classification and the misclassified label information with a greater inclination towards the accurate class information, we continue to employ the cross-entropy function for loss calculation under hard labels. The cross-entropy loss function constructed with hard labels is defined as:

$$CEL(y, \hat{y}) = -\sum_{i} y_i \ln \hat{y}_i, \tag{10}$$

where y stands for the ground truth and $\hat{y}$ the predicted value. The advantage of utilizing hard label distribution (one-hot encoding) lies in its explicit specification of the class for each sample, providing a clear error signal. Additionally, the computation of gradients for parameters in cross-entropy loss is relatively straightforward, facilitating more efficient parameter updates during the backpropagation process. This, in turn, aids the network in converging to the optimal solution more rapidly.

However, the cross-entropy loss constructed from hard labels excels in learning inter-class information, given its utilization of inter-class competition mechanism. It only focuses on the accuracy of predicting the correct label probability, disregarding distinctions among other non-correct labels. This can lead to the learned features being relatively scattered. Consequently, the network may become overly absolute in its predictions during training, which potentially diminish the generalization ability of the model. Moreover, large-scale datasets often contain mislabeled data, implying that neural networks should inherently approach the "truth" with a degree of skepticism to mitigate modeling around erroneous answers in extreme cases.

In contrast to hard label distributions, soft labels leverage the information from neighboring head pose classes to a fuller extent. In order to better capture the surrounding neighbor instance, the Kullback–Leibler (KL) divergence is employed to measure the distance between the real head pose label distribution $\mathbf{D}_i$ and the predicted head pose label distribution $\widehat{\mathbf{D}}_i$. It is defined as follows:

$$\begin{aligned} KL\left(\mathbf{D}_i \parallel \widehat{\mathbf{D}}_i\right) &= \sum_{j} \mathbf{D}^{j}_{i} \ln \frac{\mathbf{D}^{j}_{i}}{\widehat{\mathbf{D}}^{j}_{i}} \\ &= \sum_{i} \sum_{j} d^{y_j}_{X_i} \ln \frac{d^{y_j}_{X_i}}{p\left(y_j \mid X_i; \theta\right)}, \end{aligned} \tag{11}$$

where $\widehat{\mathbf{D}}_i$ means the predicted pose label distribution generated by a set of head pose images trained by the model, $\mathbf{D}_i$ denotes the pseudo-real pose label distribution we constructed. The learned mapping function $p(y|X)$ of DADL model is assumed that a parameters model $p(y|X;\theta)$ learned from E, where θ is the vector of model parameters.

Considering that by using bin classification, we use a very stable softmax layer, cross-entropy, and KL divergence. Consequently, it allows the network to predict the head pose more robustly and obtain an accurate category prediction. However, employing only the coarse classification module for the head pose task does not allow the model to obtain a more accurate angle value, so to obtain fine-grained

predictions, we added a mean square error function to the network to improve the head pose prediction. Indeed to obtain fine-grained predictions, we sum the features normalized by softmax as probabilities and the corresponding categories by multiplying them to calculate the expectation, which is summed with the ground truth angle to calculate the MSE loss. The mean square error is defined as:

$$MSE(y,\hat{y}) = \frac{1}{n}\sum_{i=0}^{n}\left(y_i - \hat{y}_i\right)^2, \tag{12}$$

where y and $\hat{y}$ respectively designate the ground truth and the predicted value. Eventually, combining the head pose label distribution learning module with cross entropy and the mean square error loss function, the global loss function is:

$$\begin{aligned} L =& CEL(y,\hat{y}) + \alpha \cdot MSE(y,\hat{y}) \\ &+ \beta \cdot KL\left(\mathbf{D}_i \parallel \hat{\mathbf{D}}_i\right), \end{aligned} \tag{13}$$

where y and $\hat{y}$ respectively designate the ground truth and the predicted value, α and β are the regression coefficients.

3.6. Optimization

Most convolutional neural networks for head pose estimation models solely utilize cross-entropy or mean square error loss functions to penalize their networks and thus predict the rotation angle of the 3D head pose. In contrast, the loss function calculated using the cross-entropy model with hard labels only considers the loss of the correct label position and ignores the loss of other positions, which leads to ignoring the loss of other incorrect label positions at the same time. Consequently, the model places enormous emphasis on increasing the probability of predicting correct labels while neglecting to reduce the probability of predicting incorrect labels, eventually leading to overfitting of the model.

Therefore, we take advantage of head pose label distribution learning. The smooth label design leads to a loss of the model, considering not only the loss of correct label positions in the training sample but also the loss of other incorrect label positions and, eventually, forcing the model to optimize in the direction of increasing the probability of correct classification and decreasing the probability of incorrect classification. In this way, the generalization ability of the model can be improved. Based on the above analysis, the best model parameter vector θ^* is:

$$\begin{aligned} \theta^* &= \arg\min_{\theta} \sum_i \sum_j d_{X_i}^{y_j} \ln \frac{d_{X_i}^{y_j}}{p\left(y_j \mid X_i;\theta\right)} \\ &= \arg\min_{\theta} \sum_i \sum_j d_{X_i}^{y_j} \ln d_{X_i}^{y_j} - d_{X_i}^{y_j} \ln p\left(y_j \mid X_i;\theta\right) \\ &= -\arg\min_{\theta} \sum_i \sum_j d_{X_i}^{y_j} \ln p\left(y_j \mid X_i;\theta\right), \end{aligned} \tag{14}$$

The softmax function is used to calculate the probability of a sample as follows:

$$p\left(y_j \mid X_i;\theta\right) = \frac{\exp\left(R\left(\theta_j X_i\right)\right)}{\sum_k \exp\left(R\left(\theta_k X_i\right)\right)}, \tag{15}$$

where θ denotes the parameters of the network model and R(·) is the output of the pre-trained ResNet network model.

Eventually, combining the head pose label distribution learning module with cross entropy and the mean square error loss function, the global loss function is:

$$\begin{aligned} L(\theta) =& CEL(y,\hat{y}) + \alpha \cdot MSE(y,\hat{y}) \\ &+ \beta \cdot (-\arg\min_{\theta} \sum_i \sum_j d_{x_i}^{y_j} \ln p(y_j \mid X_i;\theta)) \\ =& -(1+\alpha) \sum_i \sum_j d_{X_i}^{y_j} \ln p(y_j|X_i;\theta) \\ &+ \beta \cdot \frac{1}{n} \sum_i \sum_j \left(d_{X_i}^{y_j} - p(y_j|X_i;\theta)\right)^2, \end{aligned} \tag{16}$$

Fig. 6. Sample images of the three datasets. The first, second and third rows are samples from the 300W-LP, AFLW2000 and BIWI datasets, respectively.

where y and $\hat{y}$ respectively designate the ground truth and the predicted value, α and β are the regression coefficients, and θ denotes the parameters of the network model.

The mini-batch gradient descent (MBGD) method is adopted to minimize the objective function . The rule for the optimization iteration of the parameter is:

$$\theta_j \leftarrow \theta_j - \eta \frac{\delta L(\theta)}{\delta \theta_j}, \tag{17}$$

where η is the learning rate. According to the chain derivation rule is able to iteratively compute from layer to the first layer. The partial derivatives are computed for the parameter as follows:

$$\begin{aligned} \frac{\delta L(\theta)}{\delta \theta_j} =& -(1+\alpha) \sum_i \sum_j d_{X_i}^{y_j} \frac{1}{p(y_j|X_i;\theta)} \frac{\delta p(y_j|X_i;\theta)}{\delta \theta_j} \\ &+ \beta \cdot \frac{2}{n} \Sigma_i \Sigma_j d_{X_i}^{y_j} - p(y_j|X_i;\theta) \frac{\delta p\left(y_j|X_i;\theta\right)}{\delta \theta_j} \end{aligned} \tag{18}$$

Eventually, updating θ_j by Eq. (17) after $\frac{\delta L(\theta)}{\delta \theta_j}$ is obtained.

4. Experiments

4.1. Datasets

To perform a valuable and comprehensive evaluation of our network, we used three popular datasets (BIWI (Zhu and Ramanan, 2012), 300W-LP (Sagonas et al., 2013) and AFLW2000 (Zhu et al., 2016)) to train and evaluate our model. A few samples of datasets are shown in Fig. 6. These datasets were collected in different ways, from restricted laboratory settings to unconstrained field environments.

BIWI dataset is composed of videos of the Kinect device recording different head pose on 20 different subjects in a laboratory setting, and the total number of frames in this dataset is 15,000. The three angles vary in direction: pitch is in the range of −60° to +60°, yaw is between −75° and +75°, and roll is in the range of −50° to +50°.

300W-LP dataset is a large pose dataset obtained by deformation flipping on the 300 W dataset, which has more than 120,000 images containing 3837 volunteers, each with 68 key point labels and three head pose angle labels.

AFLW2000 is a challenging dataset, acquired and annotated in the field, collected from the Internet under completely unconstrained conditions. Both frontal and lateral images were collected under nine different lighting conditions, and the dataset contains precise fine-grained pose annotations that were used primarily as a test set for our purposes.

4.2. Evaluation protocols

To be able to evaluate our proposed model on different datasets fully, our experiments were performed according to the following two commonly used protocols, which are the same as HopeNet (Ruiz et al., 2018).

4.2.1. Protocol 1

Since HopeNet performs head pose estimation without landmark annotation, we follow the HopeNet setup: training on a synthetic 300W-LP dataset and testing on two real datasets, BIWI and AFLW2000. As with the HopeNet setup, when testing on both datasets, BIWI and AFLW2000, we cropped based on face landmark labels in the area around the head and did not use tracking.

4.2.2. Protocol 2

In this protocol, we take 70% of all video frames in the BIWI dataset video as the training set and the remaining 30% as the test set. And in this protocol, the video frames include different modes, including RGB, depth and time, and our proposed method is performed only under RGB frames. For a fair comparison, our protocols all follow the preprocessing strategy of the other methods, keeping only samples with Euler angles between −99° and 99°.

4.3. Implementation

In this paper, we use ResNet-50 pretrained on the 300W-LP database as the backbone. For all the databases, the face in each image is detected and cropped according to the eye positions. Then, the facial image is resized to the size of 256×256. During training, all images were first normalized using ImageNet (Deng et al., 2009) mean and standard deviation and then randomly cropped to a size of 224×224. All our models are trained on a single NVIDIA RTX 8000 GPU using PyTorch, with a batch size of 32 for 300W-LP and the other HPE databases. We divided the images with head pose image angles within ±99° into a total of 66 categories according to 3° a category, where n=66. We set the variance of the model for training on the AFLW2000 and BIWI datasets to $\sigma_y = 1$, $\sigma_p = 2$, $\sigma_r = 0.5$ and $\sigma_y = 0.35$, $\sigma_p = 0.6$, $\sigma_r = 0.3$, respectively. Where $\mu_y = \mu_p = \mu_r = 0$. Regarding the model's hyperparameters, we set the maximum limit of the batch size to 128 due to the GPU memory limitation. In our experiments, the learning rate is 1e−5, and the momentum and weight decay are set to 0.9 and 1e−4, respectively.

4.4. Competing methods

To demonstrate the effectiveness of our method, we compare our method with the following state-of-the-art methods for head pose estimation. The first group of methods is based on landmarks. Dlib, (Kazemi and Sullivan, 2014) containing landmark detection, face detection, and various other techniques, employs facial landmarks to estimate head pose. FAN (Bulat and Tzimiropoulos, 2017) method obtains multiscale information for pose estimation by merging feature blocks multiple times across layers. 3DDFA (Zhu et al., 2016) uses neural network to fit 3D models to generate head pose angles. (Kumar et al., 2017) utilizes a modified GoogleNet (Szegedy et al., 2015) architecture to predict facial points simultaneously and head pose using a multi-task strategy.

The next group of approaches utilized a landmark-free estimation. HopeNet (Ruiz et al., 2018) uses a fine-grained strategy to construct the Euler angular loss, employing ResNet (He et al., 2016b) as the backbone network and training and testing it using MSE and cross-entropy. (Yang et al., 2019) also proposes a soft-stage regression approach called FSA-Net and achieved good performance in combination with feature aggregation. (Huang et al., 2020) proposes a novel method named HPE-40 using two-stage ensembles for head pose estimation. (Zhang et al., 2020b) proposes the FDN, employing a feature decoupling approach to obtain proprietary features from different angles for estimating head pose. SCR-AT (Zhang et al., 2020a) is a approach with Gaussian distribution and spatial attention structure to predict head pose angles.

Alternative approaches in different multiple modalities are also compared. DeepHeadPose (Mukherjee and Robertson, 2015) uses both classification and regression strategy to regress the head pose directly from the RGB-D image. (Gu et al., 2017) employed VGG16 and VGG16+RNN combined with Bayesian filters to predict head pose from RGB images and RGB+Time images, respectively. (Martin et al., 2014) estimates the head pose from the depth image by building a 3D head model.

Table 1
Comparison of other state-of-the-art methods on the AFLW2000 dataset. All methods were trained on the 300W-LP dataset.

Methods	Yaw	Pitch	Roll	MAE
Dlib	23.1	13.6	10.5	15.8
3DDFA	5.4	8.53	8.25	7.39
FAN	6.36	12.3	8.71	9.12
HopeNet($\alpha = 1$)	6.92	6.64	5.67	6.41
HopeNet($\alpha = 2$)	6.47	6.56	5.44	6.16
FSA-Net	4.50	6.08	4.64	5.07
HPE-40	4.87	6.18	4.80	5.28
SSR-Net-MD	5.14	7.09	5.89	6.01
WHENet	5.11	6.24	4.92	5.42
DADL	3.99	5.82	4.21	**4.67**

Table 2
Comparison of other state-of-the-art methods on the BIWI dataset. All methods were trained on the 300W-LP dataset.

Methods	Yaw	Pitch	Roll	MAE
Dlib	16.8	13.8	6.19	12.2
3DDFA	36.2	12.3	8.78	19.1
FAN	8.53	7.48	7.36	7.89
HopeNet($\alpha = 1$)	5.17	6.98	3.39	5.18
HopeNet($\alpha = 2$)	8.80	17.3	16.2	13.9
FSA-Net	4.81	6.61	3.27	4.90
HPE-40	4.27	4.96	2.76	4.00
SSR-Net-MD	4.57	5.18	3.12	4.29
WHENet	4.49	6.31	3.61	4.65
DADL	4.18	6.19	3.27	**4.54**

4.5. Competing experiment

4.5.1. Results with protocol 1

In this case, our model is trained on the 300W-LP dataset and then tested on the AFLW2000 and BIWI datasets. The performance of our method outperforms the above methods. Tables 1 and 2 compare our model with other state-of-the-art methods on the AFLW2000 and BIWI datasets, respectively. The mean absolute error (MAE) was used as an evaluation metric. The features of the training and test datasets in this protocol are completely different; the training dataset 300W-LP is synthetic, while the test dataset BIWI is real.

Table 1 shows the results evaluated on the AFLW2000 dataset, where the proposed DADL always exhibits the smallest errors in yaw and roll. The MAE of DADL is lowest when the weights $\alpha = 2$ and $\beta = 1$ are based on the anisotropy of soft label loss and hard label loss, and the MAE decreases to 4.69. In fact, our network considers each angle of head pose estimation separately, and for three different angles the network is able to extract anisotropic feature information, making the model have better generalization ability and accurate performance.

Table 2 reports the performance of the proposed method on the BIWI dataset. Since the BIWI dataset was collected in a laboratory setting, it has less noise leading to lower MAE than the results of other datasets. Our proposed DADL employs an anisotropy-based soft label distribution to model each angle separately, which can further constrain the model and predict the corresponding attitude angle more accurately.

4.5.2. Results with protocol 2

This protocol follows the FSA-Net setting, where BIWI is utilized as both a training and test set. And the BIWI dataset is randomly split in the ratio of 7:3 for training and testing, respectively. As shown in Table 3, the performance of the approach adopting different modes on the BIWI dataset is demonstrated, with the RGB-based groups using only a

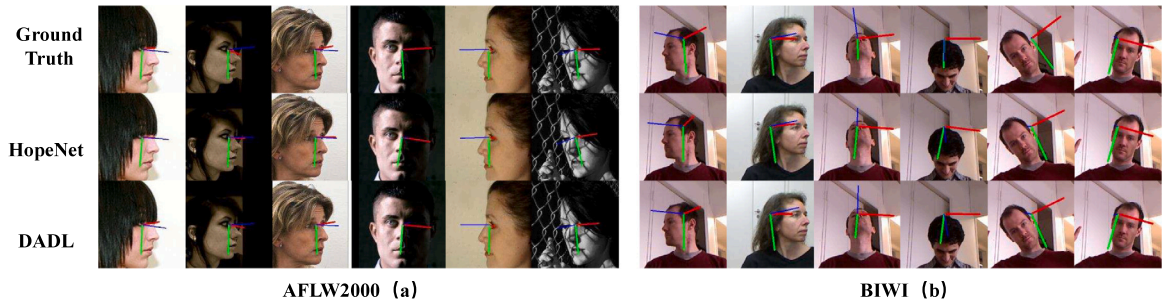

Fig. 7. Example images with converted Euler angle visualization from the AFLW2000 (a) and BIWI (b) dataset.

Table 3
Comparison with other state-of-the-art methods on BIWI datasets. 70% of the BIWI datasets were used for training and the others used for testing. Different types of methods support different modes of the dataset, including RGB, RGB + DEPTH, and RGB + TIME.

Methods	Yaw	Pitch	Roll	MAE
RGB				
DeepHeadPose	5.67	5.18	–	–
SSR-Net-MD	4.24	4.35	4.19	4.26
VGG16	3.91	4.03	3.03	3.66
FSA-Net	2.89	4.29	3.60	3.60
DADL	3.38	3.46	3.12	3.32
RGB+Depth				
DeepHeadPose	5.23	4.76	–	–
Martin	3.60	2.50	2.60	2.90
RGB+Time				
VGG16+RNN	3.14	3.48	2.60	3.07

single RGB image, while RGB+Depth and RGB+Time additionally use depth and time information, respectively. Our proposed DADL method uses only a single RGB image and outperforms RGB-based groups.

4.6. Visualization

In this section, we compare the performance of our proposed DADL model with HopeNet by taking random samples from the AFLW2000 and BIWI datasets. A visualization of the performance of our method is shown in Fig. 7. As depicted in Fig. 7, when subjecting the sample to various distortions like occlusion, illumination changes, and extreme poses, the DADL model predicts angles that align more closely with the ground truth. Different approaches estimate the head pose of these disturbed images to detect the robustness of the model. Evidently, our approach not only demonstrates superior accuracy in predicting standard images but also sustains heightened robustness in estimating extreme head poses. In contrast, when it comes to estimating extreme head poses, alternative methods falter in effectiveness due to their failure to consider the anisotropic nature of head pose.

For a more intuitive assessment of the prediction performance of both DADL and HopeNet, this paper provides visualizations in Fig. 8, showcasing predicted values and ground truth derived from samples sourced from the AFLW2000 dataset. In the visualization, small cubes represent the predicted values, large cubes represent the ground truth, the yellow line represents the ground truth direction, and the green, blue, and yellow lines represent the predicted directions. Fig. 8(a) showcases the experimental results of the HopeNet model on AFLW2000 samples, while Fig. 8(b) illustrates the experimental results of DADL on AFLW2000 samples. By comparing the experimental results, it can be observed that the angle between the predicted direction of DADL and the ground truth direction is smaller than that of HopeNet.

Fig. 9 shows this paper presents visualizations of predicted and ground truth using samples extracted from the BIWI dataset. In the visualization, small cubes represent the predicted values, large cubes represent the ground truth, the yellow line represents the ground truth direction, and the green, blue, and yellow lines represent the predicted directions. Fig. 9(a) showcases the experimental results of the HopeNet model on the BIWI samples, while Fig. 9(b) illustrates the experimental results of DADL on the BIWI samples. By comparing the experimental results, it can be observed that the angle between the predicted direction of DADL and the ground truth direction is smaller than that of HopeNet.

Table 4
Performance of the different backbone structures on AFLW2000 and BIWI datasets. All methods are trained on 300W-LP dataset.

Methods	AFLW2000				BIWI			
	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
ResNet18	4.47	6.24	4.84	5.19	4.34	6.88	3.89	5.04
ResNet152	5.35	6.05	4.79	5.40	4.68	7.20	3.71	5.20
MobileNetV2	4.89	6.34	5.47	5.57	4.73	7.78	4.43	5.66
ResNet50	3.99	5.83	4.22	**4.68**	4.18	6.19	3.27	**4.54**

Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al., 2017) is a technique employed for interpreting and visualizing the prediction outcomes of deep neural networks. Its purpose is to aid in comprehending the specific image regions and features that the model focuses on during the prediction process, thus bolstering the capacity to explain and interpret model decisions. As shown in Fig. 10, the participation regions from different methods can be observed on samples with the same identity and pose. The HopeNet method focuses more on certain facial regions, which may lead to incorrect estimation for samples with extreme pose and facial occlusions. The proposed DADL model employs an anisotropy-based soft label distribution, where adjacent head poses offer supplementary information to the target pose, thereby aiding in addressing pose estimation in these intricate scenarios. It means that DADL is able to extract more robust and detailed head pose embeddings.

4.7. Ablation studies

In this section, the performance of different backbone networks is evaluated through ablation experiments to analyze the model's total loss parameters and demonstrate the effectiveness of different components, including Gaussian-based soft label distribution modules and coarse and fine granularity modules. As shown in Table 4, ResNet18, ResNet50, ResNet152, and MobileNetV2 were selected as the backbone networks for feature extraction. These networks undergo training on the 300W-LP dataset, followed by testing on the BIWI and AFLW2000 datasets. The outcomes demonstrate that ResNet50 exhibits outstanding feature extraction capabilities, consequently yielding superior performance.

Furthermore, to verify the performance of the different components, we divided the proposed DADL model into four primary parts, w/o, Hard-label, GSoft-label and DADL module. The first segment, denoted as w/o, exclusively employs the feature extractor and fully

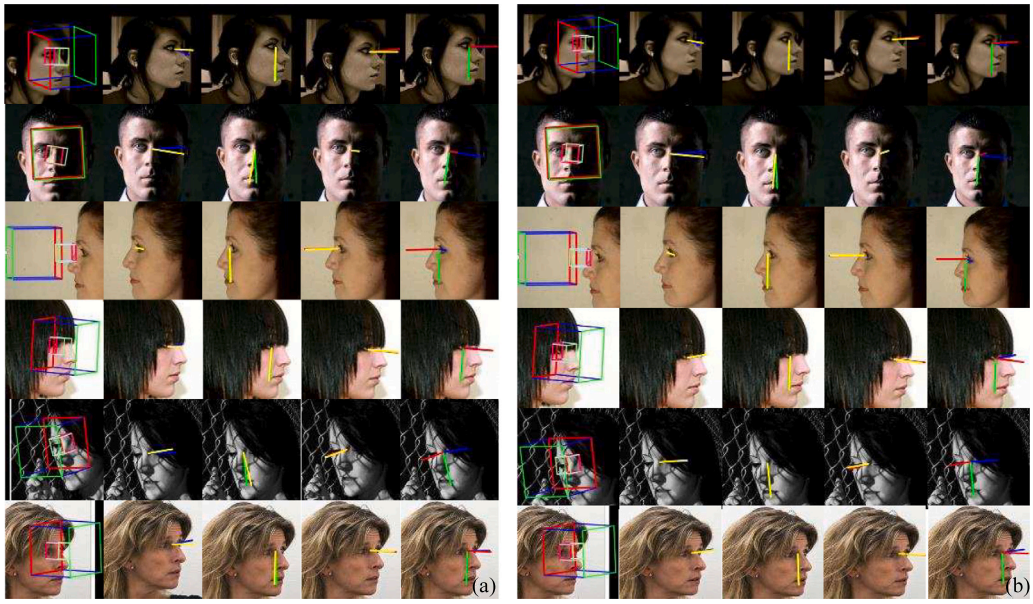

Fig. 8. (a) and (b) show the experimental results of HopeNet and ADML on the sample of AFLW2000 dataset, respectively. The first column of small cubes indicates the predicted values, the large cubes indicate the true values, the yellow line indicates the true values, and the other lines indicate the predicted values.

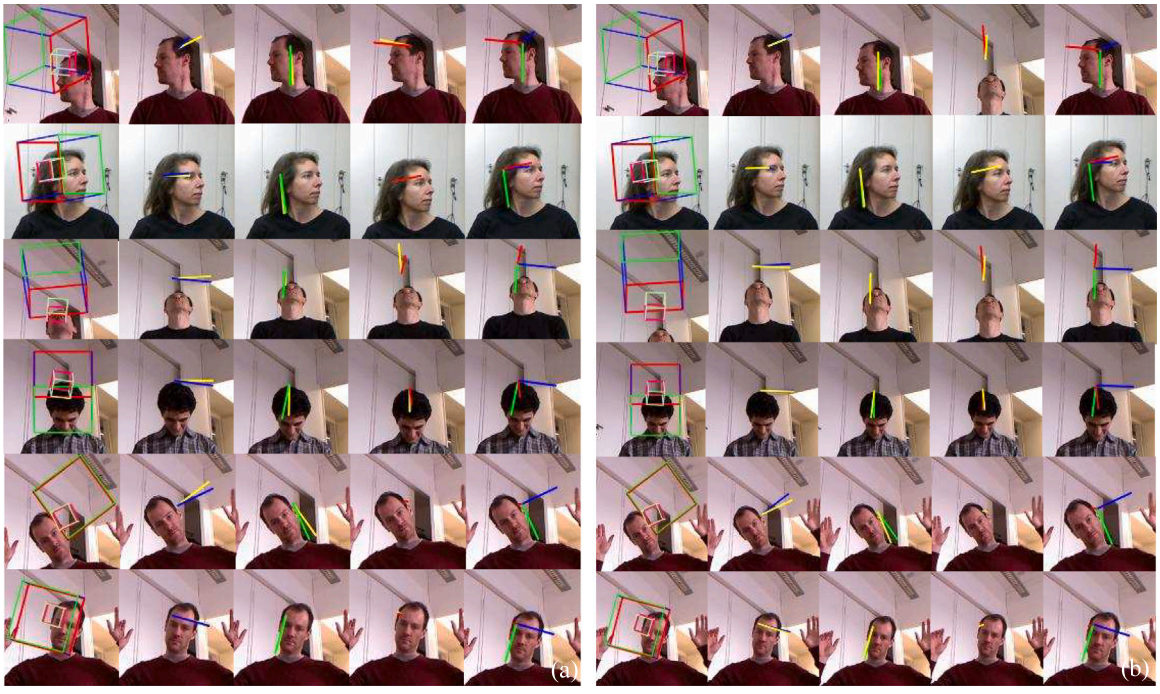

Fig. 9. (a) and (b) show the experimental results of HopeNet and ADML on the sample of BIWI dataset, respectively. The first column of small cubes indicates the predicted values, the large cubes indicate the true values, the yellow line indicates the true values, and the other lines indicate the predicted values.

connected layer without introducing any additional module. It directly applies a single Mean Squared Error (MSE) to minimize the error in output pose estimation angles. The second method adds hard labels to perform coarse-grained classification estimation first after the fully connected layer. The third part adopts a generic soft tag design instead of hard tags for pose estimation and does not consider the head pose anisotropy property. The last part of the DADL module employs MSE for finer-grained regression in a coarse-to-fine manner while adding an anisotropy-based soft label design. Then it calculates model loss utilizing KL divergence and cross-entropy. For all the above sections, the feature extractor is ResNet50. These segments are trained on the 300W-LP dataset and subsequently evaluated on the AFLW2000 and BIWI datasets. As shown in Table 5, the DADL approach has the minimum error on both datasets. It is evident that the model attains a heightened level of robustness and achieves optimal results when all modules are integrated.

Table 5
Ablation experiments for different components on the AFLW2000 and BIWI datasets.

Methods	AFLW2000	BIWI
w/o	5.56	5.75
Hard-label	5.32	5.29
GSoft-label	5.08	4.94
DADL	**4.67**	**4.54**

Fig. 10. Comparison of Grad-CAM generated by different methods on multiple samples with the same identity and different pose, indicate the experimental effects in AFLW2000 and BIWI, respectively.

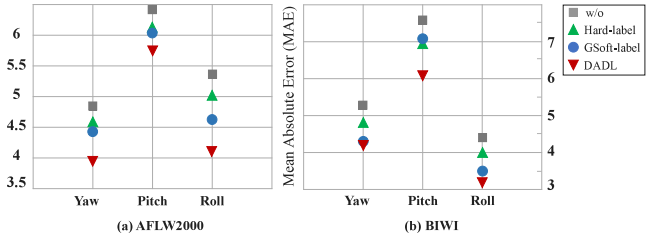

Fig. 11. Comparison of MAE of different components at different angles under protocol I.

Fig. 11 shows a detailed comparison of the yaw, pitch, and roll for several settings. Detailed information on the experimental results of each module at different angles is illustrated.

4.8. Hyperparameter experiment

In this section, the parameters associated with the total losses are scrutinized through hyperparametric experiments. The total loss is computed by summing the KL divergence, cross-entropy loss, and MSE loss, each weighted differently. As shown in Equation (16) (total loss function L), the total loss consisting of different weights significantly impacts the model performance. Therefore, we utilize the 300W-LP dataset as the training set, and to measure the generalization ability of the model, we adopt different weights on the AFLW2000 dataset and the BIWI dataset to test the model. The performance comparison employing different weights is shown in Fig. 12. It can be observed that the DADL model achieves optimal performance for total loss weighting $\alpha = 3$ and $\beta = 3$ on the AFLW2000 dataset and total loss weighting $\alpha = 1$ and $\beta = 2$ on the BIWI dataset, respectively.

5. Conclusion

In this study, we introduce a novel approach called Double Asymmetric Distribution Learning (DADL) for head pose estimation. The main objective of this research is to propose a head pose estimation method based on dual asymmetry attributes and introduces a label distribution learning mechanism to enhance the performance of the model.

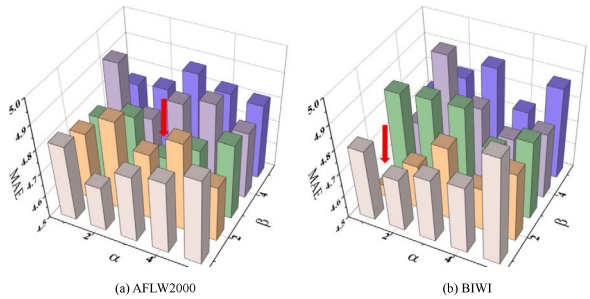

Fig. 12. Comparison of MAE of DADL using total losses with different α and β. All models are trained on 300W-LP train sets, then evaluated on 300W-LP validation sets, AFLW2000 and BIWI datasets respectively.

This allows the model to consider the information between neighboring poses of the target image while learning the ground truth of the image. Experimental results on several popular benchmarks demonstrate that our method is effective. In future work, we will further explore how to remove irrelevant features related to head pose tasks in order to construct more accurate label distributions and improve the precision of the model. This feature denoising task is an important research direction in the field of head pose estimation, and its results will contribute to improving the interpretability and generalization ability of the model.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by Major Project of the National Social Science Fund (No. 21&ZD334), National Key R&D Program of China (No. 2022YFF0901902), and the Open Research Fund of Intellectual Computing Laboratory for Cultural Heritage of Wuhan University (No. 2023ICLCH009).

References

Abdullah, D., Murtza, I., Khan, A., 2013. Feature extraction and reduction strategy based on pyramid HOG and hierarchal exploitation of cortex-like mechanisms. In: INMIC. IEEE, pp. 160–165.

Banfi, F., Pontisso, M., Paolillo, F.R., Roascio, S., Spallino, C., Stanga, C., 2023. Interactive and immersive digital representation for virtual museum: VR and AR for semantic enrichment of museo nazionale romano, antiquarium di lucrezia romana and antiquarium di villa dei quintili. ISPRS Int. J. Geo-Inf. 12 (2), 28.

Borghi, G., Venturelli, M., Vezzani, R., Cucchiara, R., 2017. POSEidon: Face-from-depth for driver pose estimation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, pp. 5494–5503.

Bulat, A., Tzimiropoulos, G., 2017. 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. pp. 1021–1030.

Cao, Z., Chu, Z., Liu, D., Chen, Y.V., 2021. A vector-based representation to enhance head pose estimation. In: IEEE Winter Conference on Applications of Computer Vision. WACV 2021, Waikoloa, HI, USA, January 3-8, 2021, IEEE, pp. 1187–1196.

Chen, J., Chen, D., Li, X., Zhang, K., 2014. Towards improving social communication skills with multimodal sensory information. IEEE Trans. Ind. Inform. 10 (1), 323–330. http://dx.doi.org/10.1109/TII.2013.2271914.

Chen, J., Guo, C., Xu, R., Zhang, K., Yang, Z., Liu, H., 2022. Toward children's empathy ability analysis: Joint facial expression recognition and intensity estimation using label distribution learning. IEEE Trans. Ind. Informatics 18 (1), 16–25.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Geng, X., 2016. Label distribution learning. IEEE Trans. Knowl. Data Eng. 28 (7), 1734–1748.

Geng, X., Qian, X., Huo, Z., Zhang, Y., 2022. Head pose estimation based on multivariate label distribution. IEEE Trans. Pattern Anal. Mach. Intell. 44 (4), 1974–1991.

Gourier, N., Hall, D., Crowley, J.L., 2004. Estimating face orientation from robust detection of salient facial structures. In: FG Net Workshop on Visual Observation of Deictic Gestures, Vol. 6. Citeseer, p. 7.

Gu, J., Yang, X., De Mello, S., Kautz, J., 2017. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1548–1557.

He, K., Zhang, X., Ren, S., Sun, J., 2016a. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, pp. 770–778.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Huang, B., Chen, R., Xu, W., Zhou, Q., 2020. Improving head pose estimation using two-stage ensembles with top-k regression. Image Vis. Comput. 93, 103827.

Jia, X., Ren, T., Chen, L., Wang, J., Zhu, J., Long, X., 2019. Weakly supervised label distribution learning based on transductive matrix completion with sample correlations. Pattern Recognit. Lett. 125, 453–462.

Kazemi, V., Sullivan, J., 2014. One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1867–1874.

Kumar, A., Alavi, A., Chellappa, R., 2017. Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In: 2017 12th Ieee International Conference on Automatic Face & Gesture Recognition. Fg 2017, IEEE, pp. 258–265.

LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1 (4), 541–551.

Li, Hu, Y., Wu, X., He, R., Sun, Z., 2020. Deep label refinement for age estimation. Pattern Recognit. 100, 107178.

Li, Zhang, D., Li, M., Lee, D., 2023a. Accurate head pose estimation using image rectification and a lightweight convolutional neural network. IEEE Trans. Multimedia 25, 2239–2251.

Li, Z., Zhang, Q., Zhu, F., Li, D., Zheng, C., Zhang, Y., 2023b. Knowledge graph representation learning with simplifying hierarchical feature propagation. Inf. Process. Manage. 60 (4), 103348.

Liu, T., Wang, J., Yang, B., Wang, X., 2021. NGDNet: Nonuniform Gaussian-label distribution learning for infrared head pose estimation and on-task behavior understanding in the classroom. Neurocomputing 436, 210–220.

Martin, M., Van De Camp, F., Stiefelhagen, R., 2014. Real time head model creation and head pose estimation on consumer depth cameras. In: 2014 2nd International Conference on 3D Vision, Vol. 1. IEEE, pp. 641–648.

Mukherjee, S.S., Robertson, N.M., 2015. Deep head pose: Gaze-direction estimation in multimodal video. IEEE Trans. Multimed. 17 (11), 2094–2107.

Murtza, I., Abdullah, D., Khan, A., Arif, M., Mirza, S.M., 2017. Cortex-inspired multilayer hierarchy based object detection system using PHOG descriptors and ensemble classification. Vis. Comput. 33, 99–112.

Murtza, I., Khan, A., Akhtar, N., 2019. Object detection using hybridization of static and dynamic feature spaces and its exploitation by ensemble classification. Neural Comput. Appl. 31, 347–361.

Ranjan, R., Patel, V.M., Chellappa, R., 2017. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans. Pattern Anal. Mach. Intell. 41 (1), 121–135.

Ruiz, N., Chong, E., Rehg, J.M., 2018. Fine-grained head pose estimation without keypoints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 2074–2083.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2013. 300 Faces in-the-wild challenge: The first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 397–403.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Song, C., Wang, S., Chen, M., Li, H., Jia, F., Zhao, Y., 2023. A multimodal discrimination method for the response to name behavior of autistic children based on human pose tracking and head pose estimation. Displays 102360.

Sun, Y., Wang, X., Tang, X., 2013. Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3476–3483.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.

Thai, C., Tran, V., Bui, M., Nguyen, D., Ninh, H., Tran, H., 2022. Real-time masked face classification and head pose estimation for RGB facial image via knowledge distillation. Inform. Sci. 616, 330–347.

Wang, J., Zhang, Q., Shi, F., Li, D., Cai, Y., Wang, J., Li, B., Wang, X., Zhang, Z., Zheng, C., 2023a. Knowledge graph embedding model with attention-based high-low level features interaction convolutional network. Inf. Process. Manage. 60 (4), 103350.

Wang, S., Zhao, A., Lai, C., Zhang, Q., Li, D., Gao, Y., Dong, L., Wang, X., 2023b. GCANet: Geometry cues-aware facial expression recognition based on graph convolutional networks. J. King Saud Univ. - Comput. Inf. Sci. 35 (7), 101605.

Wen, X., Li, B., Guo, H., Liu, Z., Hu, G., Tang, M., Wang, J., 2020. Adaptive variance based label distribution learning for facial age estimation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (Eds.), Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII. In: Lecture Notes in Computer Science, vol. 12368, Springer, pp. 379–395.

Xu, Dong, Y., Wu, J., Sun, Z., Shi, Z., Yu, J., Gao, S., 2018. Gaze prediction in dynamic 360 immersive videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5333–5342.

Xu, Gu, S., Tao, H., Hou, C., 2021. Fragmentary label distribution learning via graph regularized maximum entropy criteria. Pattern Recognit. Lett. 145, 147–156.

Yang, T.-Y., Chen, Y.-T., Lin, Y.-Y., Chuang, Y.-Y., 2019. 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. pp. 1087–1096.

Zaremba, W., Sutskever, I., Vinyals, O., 2014. Recurrent neural network regularization. arXiv preprint arXiv:1409.2329.

Zhang, Fu, K., Wang, J., Cheng, P., 2020a. Learning from discrete Gaussian label distribution and spatial channel-aware residual attention for head pose estimation. Neurocomputing 407, 259–269.

Zhang, Wang, M., Liu, Y., Yuan, Y., 2020b. FDN: Feature decoupling network for head pose estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 07. pp. 12789–12796.

Zhao, Z., Liu, Q., Zhou, F., 2021. Robust lightweight facial expression recognition network with label distribution training. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 4. pp. 3510–3519.

Zhou, Y., Gregson, J., 2020. Whenet: Real-time fine-grained estimation for wide range head pose. arXiv preprint arXiv:2005.10353.

Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z., 2016. Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 146–155.

Zhu, X., Ramanan, D., 2012. Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2879–2886.