Pattern Recognition 164 (2025) 111536

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

ADGaze: Anisotropic Gaussian Label Distribution Learning for fine-grained gaze estimation

Duantengchuan Li [a,b], Shutong Wang [a], Wanli Zhao [a,b,*], Lingyun Kang [c,d], Liangshan Dong [e], Jiazhang Wang [f], Xiaoguang Wang [a,b]

[a] School of Information Management, Wuhan University, Wuhan 430072, China
[b] Intelligent Computing Laboratory for Cultural Heritage, Wuhan University, Wuhan 430072, China
[c] School of psychology, Jiangxi Normal University, Nanchang 330022, China
[d] Department of Educational Technology, Jiangxi Normal University, Nanchang 330022, China
[e] School of Physical Education, China University of Geosciences, Wuhan 430074, China
[f] Department of Eletrical and Computer Engineering, Northwestern University, Evanston, IL 60208, USA

ARTICLE INFO

Keywords:
Gaze estimation
Anisotropic characteristic
Label distribution learning
Deep learning

ABSTRACT

Gaze estimation technology is crucial for enhancing the effectiveness and safety of applications in human–computer interaction, intelligent driving, virtual reality, and medical diagnosis. With advancements in deep learning, gaze estimation methods using deep neural networks have been extensively researched and applied. However, existing methods have yet to address the anisotropic characteristics of eye features. Based on the discovered anisotropic characteristics, we propose an Anisotropic Gaussian Label Distribution Learning Network for Gaze Estimation (ADGaze). ADGaze is capable of catching neighboring information by taking advantage of coarse-to-fine methodology and the anisotropic soft label construct. The coarse-to-fine framework initially performs classification tasks for gaze estimation, grouping gaze images with small variations into the same category, followed by regression tasks for each category. The construction of anisotropic Gaussian label distributions adopts methods based on data statistics and feature similarity. Extensive experimentation on public datasets has been carried out to substantiate the efficacy of this model. Our code is publicly available at https://github.com/dacilab/ADGaze.

1. Introduction

Gaze estimation (GE) has increasingly attracted scholarly interest as a crucial technology in human–computer interaction and virtual reality in recent years [1]. Initial methodologies predominantly relied on model-based gaze estimation strategies [2,3] focusing on 2D or 3D modeling of the eye based on critical features like corneal reflections, the pupil's center, the Purkinje image's center and iris outlines, leveraging these biological characteristics to ascertain the gaze orientation. While model-based methods can furnish precise gaze angles, they require auxiliary sensor equipment to attain substantial estimation accuracy, leading to practical operational distances being restricted by the limits of the hardware's reach. Due to this constraint, researchers have introduced feature-learning-driven gaze estimation methodologies, principally divided into geometry-based [4,5] and appearance-based methods [6,7]. Geometry-based methods notably offer a significant decrease in parameter calibration; nonetheless, their effectiveness is largely contingent upon the amount of training data and the model's potency. Conversely, appearance-based methods have continually faced challenges in improving accuracy, which could be enhanced by the scarcity of datasets in gaze estimation and the constraints of contemporary feature extraction methodologies. In gaze estimation technologies, model-based approaches aim to precisely simulate individual eye structures. These methods are computationally intensive and usually require personalized adjustments to account for individual differences. Furthermore, they are especially sensitive to variations in lighting conditions. Geometry-based gaze estimation relies on accurately identifying and localizing key features of the eye, such as the pupil center and eye corners. While this method does not demand significant computational power, it can be limited by occlusions, blinks, and is less effective when dealing with large head pose variations. Appearance-based gaze estimation offers robust resistance to lighting variations, yet if the training data lacks adequate diversity, the models trained might capture dataset-specific features instead of generalizable gaze patterns. However, in recent years, with the increasing number

* Corresponding author at: School of Information Management, Wuhan University, Wuhan 430072, China.
E-mail addresses: dtclee1222@whu.edu.cn (D. Li), wanlizhao146@whu.edu.cn (W. Zhao).

https://doi.org/10.1016/j.patcog.2025.111536
Received 24 September 2024; Received in revised form 6 February 2025; Accepted 27 February 2025
Available online 7 March 2025
0031-3203/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

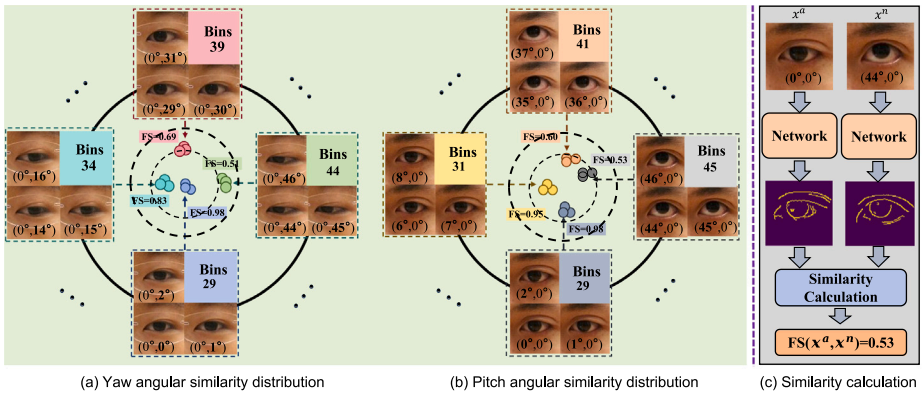


Fig. 1. The anisotropy of gaze estimation. In the Pitch and Yaw directions, the feature distributions of gaze images at various angles exhibit significant differences, a phenomenon we term as anisotropic characteristics. Specifically, we define each 3-degree interval as a Bin, within which features from different angles display high similarity and clustering. This implies that within small angular variations, gaze image features remain relatively consistent, a property that enhances classification and prediction accuracy in gaze estimation tasks.

of datasets collected in gaze estimation and the rapid development of deep learning technology in computer vision [8,9], recommender system [10–12], knowledge graph [13,14] and automatic driving [15]. The appearance-based methods [16] can use deep learning to extract more gaze features and improve the accuracy of models. However, in the in-depth study of a large number of gaze estimation datasets, we have discovered two key characteristics as follows.

The first characteristic in Fig. 1: images exhibit striking similarity when gaze angles change slightly, displaying minimal differentiation. This makes it challenging for models to directly fit and regress the gaze angles of these images, hindering the attainment of accurate estimates. Existing gaze estimation methods mainly include model-based approaches [2,3,17] and appearance-based approaches [7,18,19]. Model-based methods rely primarily on 2D or 3D modeling of the eyeball, determining the direction of gaze by analyzing key features such as corneal reflection and the center of the pupil. Although this approach can provide accurate gaze angles, it requires additional sensor equipment, limiting its application scenarios. With the development of deep learning, appearance-based methods have gradually become mainstream, the deep learning-based gaze estimation methods adopt a direct regression strategy, which struggles to extract features relevant to the gaze estimation task when the similarity of image features is high, leading to difficulty in improving the accuracy of gaze estimation. This makes it challenging for models to directly fit and regress the gaze angles of these images, hindering the attainment of accurate estimates. Accordingly, this paper proposes a coarse-to-fine methodology, initiating a classification stage for gaze estimation, clustering images with insignificant angle changes into shared categories, followed by regression tasks for each classified group. Effectively, the gaze estimation process is divided into two phases, categorization and regression, to augment the efficiency and precision of the estimation process. Partitioning the full spectrum of gaze angles into multiple categories and performing more meticulous regression assessments within these narrower bands can significantly heighten gaze estimation accuracy. Classification within these compact categories facilitates the model's identification of the target's broad range or category, diminishing the search scope for subsequent regression and enhancing the model's operational efficiency. Moreover, conventional classification tasks often utilize one-hot encoding schemes (hard labels), where each image's category is indicated by either 0 or 1. Since hard labels consider each class independently, they do not convey the similarities or distinctions among categories, hindering the model's capacity to grasp more complex category hierarchies. To address this problem, we introduce Label Distribution Learning (LDL, or soft labels) [20] into gaze estimation. LDL is an advanced machine learning approach designed to convert the single label of each sample into a probability distribution, capturing subtle differences and correlations among samples. This method demonstrates particular strengths in fields like age estimation [21] and head pose estimation [22] by enabling a finer depiction of continuous variations in data and offering better adaptability to class imbalance issues. Conventional approaches typically process the label distributions of all classes uniformly, potentially failing to capture the nuanced differences between various classes. In gaze estimation, eye features vary significantly across different angles, yet traditional LDL methods find it challenging to effectively model these variations, thus inadequately accounting for the anisotropic nature of eye features. To tackle these challenges, we utilize both data statistical information and feature similarity independently to construct label distributions. Soft labels offer probabilistic distributions across categories, signifying the likelihood of a sample falling into each class. Such probability data enriches the model with supplementary insights into the ambiguity related to the target class, ensuring that during learning, the model endeavors beyond mere rigid categorization, aiming instead to forecast a probability distribution. This facilitates the model's comprehension of inter-category resemblances, thereby bolstering its generalization performance on unseen data.

Inspired by the work of Zhao et al. [9], we found the second characteristic in Fig. 1 through further observation: gaze angles show varying feature distributions in the yaw and pitch dimensions. This is illustrated in Fig. 1(c), where a similarity computation network assesses the likeness between (0°, 0°) and the chosen images, highlighting disparities in the feature spreads of various gaze angles along both pitch and yaw orientations. We denote this observation as the anisotropic nature of gaze estimation concerning yaw and pitch angles. Consequently, crafting more tailored soft label distributions becomes essential to adequately model gaze features along divergent directional axes.

In addressing these challenges, this work presents ADGaze, a fine-grained gaze estimation method based on Anisotropic Gaussian Label Distribution Learning. Initially, by adopting the concept of classification preceding regression, the problem domain is divided into more compact subsets, thereby simplifying and enhancing the precision of subsequent regression tasks. Furthermore, we propose an anisotropic soft label distribution to learn the distinctions in gaze estimation features along vertical and horizontal orientations, constructing soft labels employing two distinct similarity metrics grounded in statistical and feature-based analyses. The primary contributions of this work are summarized as follows:

- Two essential characteristics revealed: (1) Intra-class similarity: Images with gaze angles varying within a narrow range exhibit high similarity, with minute differences between them. (2) Anisotropy of gaze angles: Distinct feature distributions exist along gaze angles' horizontal and vertical orientations.

- We propose multi-loss networks with anisotropic soft label distribution learning to model gaze estimation through a classification-first-then-regression approach. Iteratively extracting subtle features from gaze images in a coarse-to-fine manner significantly enhances the model's accuracy.
- To better capture the anisotropy of gaze estimation in pitch and yaw, we construct label distributions for different directions using statistical and feature-based similarity approaches. This enables the network to grasp richer contextual information, extracting more refined gaze estimation features.
- Experimental Validation and Results:Extensive evaluations on widely-used gaze estimation datasets, demonstrate that our proposed method outperforms existing approaches. Additionally, visualizations further affirm the efficacy of the proposed anisotropic Gaussian label distribution strategy.

The subsequent sections of this paper are organized as follows: Section 2 provides a concise review of the relevant literature on gaze estimation. Section 3 presents a detailed exposition of our proposed ADGaze model. Section 4 offers an in-depth analysis and discussion of the experimental results. Finally, Section 5 summarizes the document and draws conclusive remarks.

2. Related work

2.1. Gaze estimation

Advancements in gaze estimation technology have contributed significantly to the field of pattern recognition. A method for 3D gaze estimation without requiring explicit personal calibration was proposed, utilizing complementary gaze constraints and center priors along with learned generic facial information to achieve accurate gaze estimation while reducing dependency on annotated data specific to test subjects [23], thus enhancing generalization capabilities. To address issues of poor generalization due to data bias, another study introduced a method for appearance-debiased gaze estimation through stochastic subject-wise adversarial learning, incorporating novel loss functions and training strategies to improve the accuracy and generalization of gaze estimation [24]. The application of Transformers in cascaded learning enables simultaneous eye landmark detection, eye state detection, and gaze estimation, capturing long dependencies and implicit associations to boost performance [25]. By introducing synthetic data and consistency supervision, a gaze estimation method that employs semi-supervised eye landmark detection as an auxiliary task tackles the issue of insufficient annotations in real-world datasets [1], enhancing performance and generalization. Gaze position detection by calculating three-dimensional facial positions and motions demonstrates acceptable error rates in gaze position detection experiments [26]. In the realm of gaze stabilization within active vision, research has led to the development of phase-based vergence error extraction algorithms simplifying the matching process [27]. A gaze estimation method dependent on gazing points utilizes dynamic virtual plane projection and a coarse-to-fine framework combined with offline and online parameter learning heuristics to significantly improve gaze estimation performance [19]. However, in recent years, with the increasing number of datasets collected in gaze estimation and the rapid development of deep learning technology, appearance-based methods [16] can now leverage deep learning to extract more gaze features and improve the accuracy of models.

2.2. Appearance-based methods

In gaze estimation, appearance-based methods have proven effective by analyzing texture features in input images and learning their mapping to gaze directions. Yu et al.'s work [28] established a relationship between eye keypoint locations and gaze direction using 17 points and the pitch angle, demonstrating the potential for precise feature extraction in effective gaze estimation. Considering the influence of head pose, Wang et al. [16] introduced adversarial learning within convolutional neural network frameworks to capture variations in eye appearance and head pose, thereby improving the accuracy and generalization of gaze direction estimation. Inspired by the successful application of transformers [18] in the natural language processing field, Transformer has become a hot topic as a backbone in computer vision tasks. Appearance-based gaze estimation has also benefited from transformer-based approaches. GazeTR [6] first utilized pure transformers and hybrid transformers for gaze estimation, finding that hybrid transformers significantly outperformed pure ones with fewer parameters. Given the inefficiency of transformers in real-time edge device applications, the GazeNAS-ETH [7] model addresses gaze estimation with precision and efficiency while featuring only around 1M parameters, facilitating its deployment in real-time applications. Considering that existing eye-tracking studies mainly focus on individual tasks such as pupil detection or gaze estimation, ignoring the implicit relationships between different tasks in eye-tracking. Gou et al. [19] employ transformers to capture long-range dependencies encompassing explicit eye structure information and the implicit correlations across eye landmark detection, eye state detection, and gaze estimation tasks. However, these methods frame gaze estimation as regression, overlooking the minute eye movements and similarity in slight gaze variation images, leading to high complexity and poor data fitting. Hence, this paper innovatively suggests a coarse-to-fine strategy, categorizing small angular ranges before regression, ensuring efficient model convergence.

2.3. Label distribution learning

LDL [20], by converting single labels into a fixed distribution, offers a clear and nuanced method to describe relationships between instances, effectively capturing correlations among neighboring samples. The LDL technique has shown remarkable performance in practical applications such as age estimation and head pose estimation. Geng et al. [22] analyzed different head poses and fitted their probabilities using Gaussian-based methods and training networks with constructed soft labels, which significantly improved classification performance. In age prediction [21], the adoption of soft label distribution learning, focusing solely on sensibly contiguous ages, outperforms other current methods in accuracy. For multi-label classification tasks, Zhao et al. [29] introduced a scalable LDL approach that significantly enhances the capability to process large datasets, especially beneficial for classification issues in high-dimensional feature spaces. Regarding text classification, Zhao et al. [30] presented a variational continuous label distribution learning framework, enabling smooth adjustments in label distributions according to variations in input features, thus offering a flexible and robust solution for complex multi-label text classification tasks. In the domain of object detection, Hang et al. [31] demonstrates that employing Gaussian distributions to model label distributions enables more precise capture of object location information in spherical images, particularly excelling in handling complex backgrounds or viewpoint variations. Furthermore, in medical image segmentation, Li et al. [32] investigated a curriculum-based label distribution learning method that efficiently tackles the training challenges posed by imbalanced data by progressively adjusting the learning process according to sample difficulty levels. To our knowledge, compared to traditional deep learning approaches, LDL has significant advantages in feature capture: it seizes correlations among neighboring class labels, providing richer neighborhood information, thereby reinforcing the network's robustness, effectively addressing class imbalance, and enabling continuous prediction rather than being confined to discrete classifications. These strengths of LDL make it particularly suitable for understanding and learning the intricate relationships in human eye images. Building anisotropic label distributions allows for a more meticulous depiction of interactions and continuous changes among visual elements, enhancing

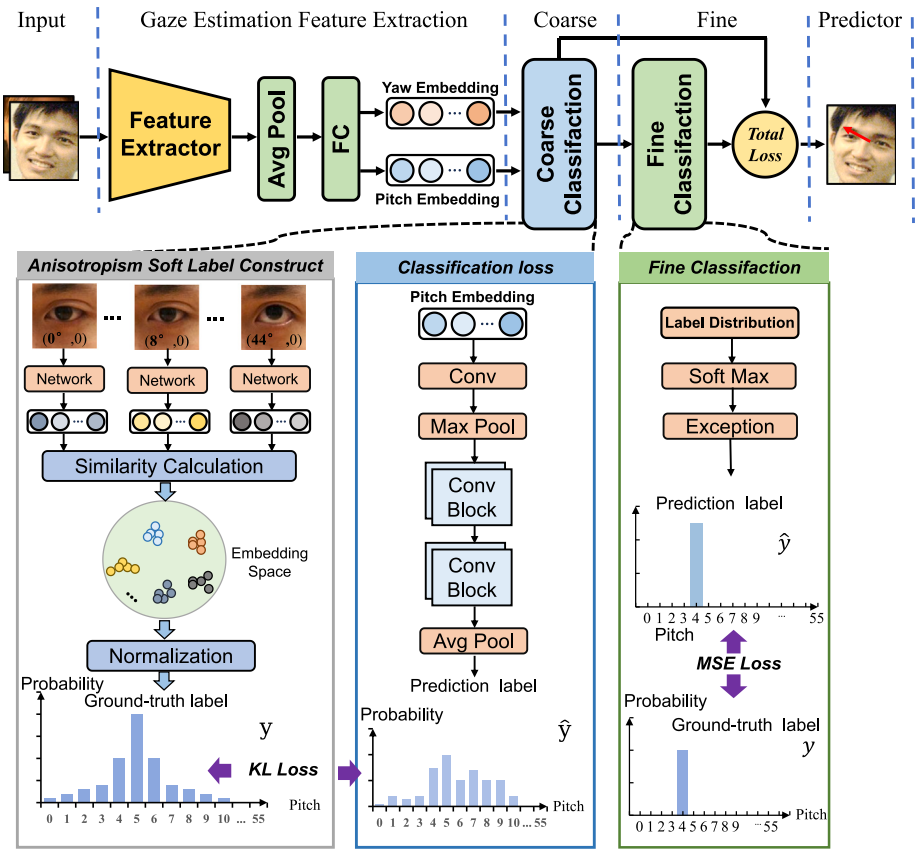


Fig. 2. Pipeline of our AFDNet model. The ADGaze model is primarily divided into three parts: Feature Extractor, Coarse Classification, and Fine Classification. Initially, facial images are fed into the Feature Extractor for gaze feature extraction. Subsequently, both classification and regression steps are utilized to compute the loss function, culminating in gaze direction prediction.

the network's capability to acquire neighborhood information. Consequently, in this work, we pioneer the integration of LDL into the task of gaze estimation and, further considering the anisotropy of gaze estimation, employ distinct distributions to fit accurate class labels, thereby narrowing the solution space of the model.

3. Proposed ADGaze model

In this section, we outline our model's overarching framework, delve into the specifics of each component, and ultimately elucidate the process by which the model is resolved.

3.1. Outline of ADGaze

The proposed ADGaze model, depicted in Fig. 2, comprises four main components: Feature Extractor, Coarse Classification, Fine Classification, and Regression. It begins with facial images being fed into the Feature Extractor, initiating the process by extracting gaze-related features. This is succeeded by a dual process of classification and regression to ascertain the loss function, concluding with the prediction of the gaze direction itself. In the feature extraction phase, ResNet50 is adopted as the primary feature extractor. Following extraction, features are bifurcated into two subsets corresponding to yaw and pitch angle characteristics, implementing a foundational design where classification precedes regression, effectively embodying a coarse-to-fine (CF) feature refinement approach. Within the Coarse stage, Cross Entropy Loss is supplanted by a distribution similarity loss function (KL), fully harnessing the characteristics of proximal labels. During the Fine phase, we estimate the expected value for every output angle per input bin and then perform a refined regression utilizing MSE. Ultimately, the KL and MSE losses are combined with consideration of their assigned weights.

3.2. Feature extractor

To address common challenges encountered during the training of deep networks and ensure high precision and stability, we opted for ResNet50 as our main feature extractor. In gaze estimation, deep learning architectures like Convolutional Neural Networks (CNNs) are paramount for extracting informative features. By scrutinizing eye images, these mechanisms distill essential cues, including eyeball placement, ocular texturing, pupil dimensions and placement, periocular skin tone, as well as aspects of head orientation and facial gestures. These collective features facilitate precise gaze direction determination. The essence of feature extraction resides in hierarchical data abstraction, going beyond the mere retrieval of rudimentary features. Following extensive testing of several backbone architectures, this work settles on ResNet50 as the chosen feature extractor. ResNet50's layer-wise extraction of local image characteristics creates a hierarchical depiction, enabling a more profound understanding of complex ocular imagery, thus augmenting the precision of gaze directional forecasts.

3.3. Coarse classification

To enhance the feature extraction capability of the model under conditions of highly similar features, the coarse classification module introduces anisotropic features into the gaze estimation model to improve classification accuracy. Focusing on the yaw angle feature as an illustrative example, the process commences in the Coarse stage with the feature passing through the Anisotropic Soft Label Construct module. Here, hard categorical assignments are transformed into probability-weighted 'truth' soft labels by leveraging Gaussian distribution principles. The compatibility of the two resultant label distributions is then quantified using KL divergence, forming the basis of

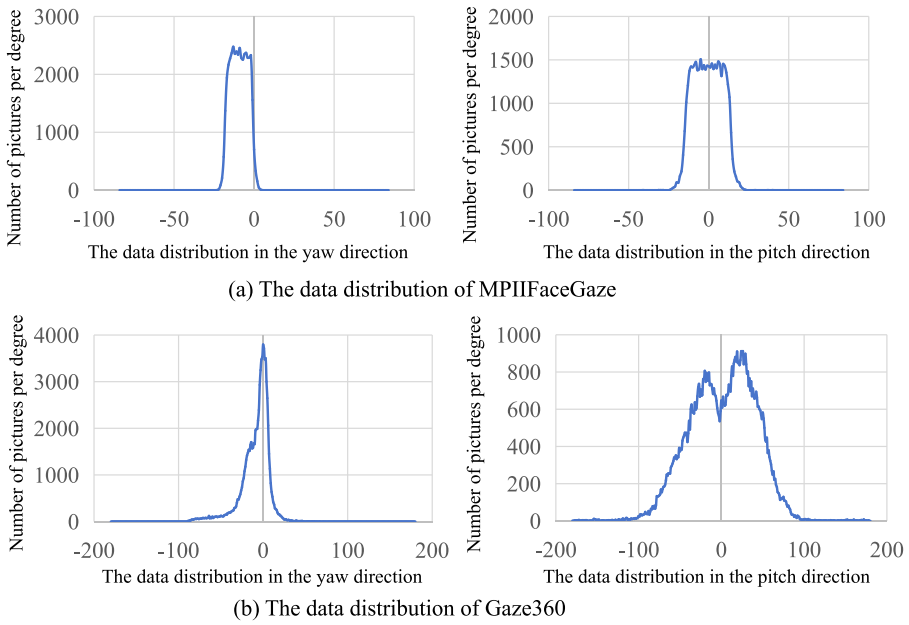


Fig. 3. The data distribution of the different datasets. The data distribution of the Gaze360 and MPIIFaceGaze datasets shows a quasi-Gaussian distribution in the Pitch and Yaw directions.

the classification loss calculation. In the context of MPIIFaceGaze, gaze direction images are systematically categorized into 56 distinct groups during the initial Coarse phase. Unique soft labels are devised for both pitch and yaw angles, drawing upon statistical proximity and feature-based similarities. Departing from the conventional SoftMax loss, KL divergence is strategically adopted to fine-tune the congruence among label distributions, amplifying the precision of classification outcomes. This meticulous design capitalizes on the intrinsic relationships among neighboring labels, thereby substantively refining the overall accuracy of classification results.

3.3.1. Anisotropic soft label construct based on data statistics

As shown in Fig. 3, the data distribution of the two most common datasets in the field of gaze estimation, Gaze360 and MPIIFaceGaze, is calculated at 1-degree intervals. The statistical range of the Gaze360 dataset spans from −180° to +180°, whereas the angle range of the MPIIFaceGaze dataset is from −84° to +84°. Statistical results reveal that the gaze angles of the Gaze360 and MPIIFaceGaze datasets are predominantly distributed within ±50° and ±20°, highlighting the extreme imbalance in gaze angle distribution, particularly for extreme angles. However, a highly imbalanced distribution of training samples can significantly reduce the accuracy of gaze estimation models. Therefore, a label distribution learning strategy is proposed to enhance the model's learning of facial images by incorporating additional knowledge from facial images at various angles, thereby improving the learning of target images. Label distribution allows the model to leverage facial images at adjacent angles, providing supplementary information when learning specific angles. This approach assigns each facial image a label distribution instead of a single true pose label. This label distribution not only aids the model in learning the true angle of the facial image but also in learning adjacent angles. This section employs two distinct Gaussian label distributions to represent facial samples in yaw and pitch angles, thereby enhancing the overall learning process. Specifically, analysis of gaze angle data distribution suggests that the gaze distribution follows a quasi-Gaussian pattern, making Gaussian distribution suitable for constructing gaze estimation label distributions. The process of mapping angle labels to the gaze estimation label distribution $\mathbf{d}$ through a mapping function constitutes the construction of the label distribution. Assume a training set $\mathbb{E} = \{(X_0, \mathbf{d}_0), (X_1, \mathbf{d}_1), \ldots, (X_n, \mathbf{d}_n)\}$, where $\mathbf{d}_i = \{q_0, q_1, \ldots, q_n\}$ is the single true label value of X_i, and n represents the batch size. In any dimension, the proportion of each angle component in a gaze estimation label distribution is represented by:

$$\mathbf{d}_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right). \tag{1}$$

The labels in the two dimensions of Yaw and Pitch can both be represented by a single formula. Here, the Gaussian label distribution is illustrated using the Yaw angle. Given a facial image X_i and a set of complete yaw angle labels $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$, if its yaw angle label is y_i, $i = 1, 2, \ldots, n$, then the corresponding yaw angle label distribution is $\mathbf{d}_{X_i}^{y_i} = \left\{d_{X_i}^{y_1}, d_{X_i}^{y_2}, \ldots, d_{X_i}^{y_n}\right\}$, where the mth dimension is as follows:

$$d_{X_i}^{y_m} = f_{X_i}^{\text{yaw}} = \frac{1}{\sqrt{2\pi}\sigma_y} exp\left(-\frac{(x - \mu_y)^2}{2\sigma_y^2}\right), \tag{2}$$

where i represents the ith bin classification of the yaw angle, y_i represents the same gaze estimation label as q_i, and μ_y and σ_y represent the mean and variance of the Gaussian function in the yaw angle, respectively. Therefore, $\mathbf{d}_{X_i}^{y_i}$ represents the extent to which the label describes the example under constraint $\sum_{k=1}^{n} \mathbf{d}_{X_i}^{y_i} = 1$, indicating that the label set y completely describes the example.

The label distribution for the other angle follows the same definition but with different parameter constraints. $\mathbf{d}_{X_i}^{p} = \left\{d_{X_i}^{p_1}, d_{X_i}^{p_2}, \ldots, d_{X_i}^{p_n}\right\}$ can be obtained through a set of pitch angle labels of X_i.

$$d_{X_i}^{p_i} = \frac{p_{X_i}^{y_i}}{\sum_{k=1}^{n} p_{X_i}^{y_i}}, \tag{3}$$

$$p_{X_i}^{y_i} = f_{X_i}^{\text{pitch}} = \frac{1}{\sqrt{2\pi}\sigma_p} exp\left(-\frac{(x - \mu_p)^2}{2\sigma_p^2}\right), \tag{4}$$

according to the distribution curves fitted by similarity, the parameters are set as variances $\sigma_y = 0.9$, $\sigma_p = 0.8$, and $\sigma_y = 0.7$, $\sigma_p = 0.2$, with the mean set as $\mu_y = \mu_p = 0$ for the MPIIFaceGaze and Gaze360 datasets, respectively.

3.3.2. Anisotropic soft label construct based on feature similarity

Considering the varying feature similarities between different categories, we further construct anisotropic Gaussian feature distributions using feature similarity calculation methods. To accurately measure the

similarity between adjacent images, we introduced the cosine function to calculate the Gaze Pose Similarity (GPS) between image features. In this process, we use a simple five-layer pre-trained network and extract the output of its last fully connected layer as the gaze features of the images. Fig. 1(c), the central image X_0 has an angle of (0°, 0°). An image is taken every 3° in the Pitch direction, with an example image at (0°, 3°) shown in Fig. 1(c). This paper uses a pre-trained five-layer network to extract the features of images X_0 and X_1, outputting the feature arrays through the last layer of the grid. The calculation process of GPS is formally described as follows:

$$\mathrm{GPS}\left(X_0, X_1\right) = \frac{FV\left(X_0\right) * FV\left(X_1\right)}{|FV\left(X_0\right)| * |FV\left(X_1\right)|}, \tag{5}$$

where FV represents the feature vector of the last layer. Eq. (5) is used to measure the similarity of images with the same angle change in the Pitch or Yaw direction.

To utilize anisotropic Gaussian feature distribution, this paper introduces Gaussian distributions for both Pitch and Yaw angles. Suppose the range of the pitch angle for the selected instance object is 0 to 55. Then, the label of an instance object can be defined as $\mathbf{d}_{X_i}^{y} = \left\{d_{X_i}^{y_1}, d_{X_i}^{y_2}, \ldots, d_{X_i}^{y_n}\right\}$, which satisfies two conditions: (1) the probability distribution should ensure that the probability value of its true label is the highest; (2) the sum of all values is 1, with each value between 0 and 1. Assuming the true Pitch category is α, its Gaussian function is defined as:

$$\mathrm{GAUS}\left(\mathbf{d}_{X_i}^{y_i} \mid \alpha; \sigma\right) = \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y^i - \alpha\right)^2}{2\sigma^2}\right)}{\sum \mathrm{GAUS}\left(\mathbf{d}_{X_i}^{y_i} \mid \alpha; \sigma\right)}. \tag{6}$$

Similarly, the Gaussian distribution for the Yaw angle can also be obtained using Eq. (6).

3.4. Fine classification

To synergize with the preceding classification tasks and achieve more accurate gaze angle predictions, a fine-grained classification and regression module is introduced. Taking the input Yaw angle feature as an example, the Yaw angle feature passes through the Coarse module and then the Fine module to obtain the predicted hard label angle value of the yaw angle. Then, the predicted hard label value and the true hard label value are measured using Mean Squared Error (MSE) loss. Finally, the Kullback–Leibler (KL) loss and MSE loss are weighted and summed together for joint optimization.

3.5. Optimization

Considering the incorporation of prior distributions in gaze estimation, our model employs Maximum A Posteriori (MAP) estimation instead of Maximum Likelihood Estimation (MLE) to identify the optimal parameters. This approach enables the model to leverage both current data and prior knowledge, thereby enhancing the accuracy and reliability of parameter estimation. Gaze estimation consists of a coarse classification component and a fine regression component. Taking the MPIIFaceGazes dataset as an example, we first construct the coarse classification component of gaze estimation. Based on the dataset, we select the range of straight-line angle changes. Assuming the selected Pitch angle in the gaze ranges from −99 to +99, with 3-degree intervals, it forms 56 bins, representing 56 categories. The classification component outputs a 66-dimensional vector, which undergoes a Softmax operation to produce a label vector that sums to 1:

$$\mathbf{d}_{X_i}^{y_i} = \mathrm{soft\,max}\left(X_i\right) = \frac{e^{X_i}}{\sum_{j=1}^{n} e^{X_j}}, \tag{7}$$

where $\mathbf{d}_{X_i}^{y_i}$ represents the probabilistic value of the ith label vector, and n represents the classification category.

Because a prior distribution is included in gaze estimation, this model should use Maximum A Posteriori (MAP) estimation instead of Maximum Likelihood Estimation (MLE) to find the parameters. Specifically, assume a set of full-face images X and their constructed one-dimensional Pitch or Yaw angle Gaussian distribution $\mathbf{d}$. MAP estimation seeks to find the set of neural network parameters that maximizes the posterior probability given X and $\mathbf{d}$. The parameters of the neural network are solved as follows:

$$\boldsymbol{\theta}^* = \arg\ \max\ M\left(X, \mathbf{d}|\boldsymbol{\theta}\right), \tag{8}$$

According to Bayes' theorem, Eq. (8) can be further written as:

$$\boldsymbol{\theta}^* = arg\ max \frac{M\left(X, \mathbf{d}|\boldsymbol{\theta}\right) \cdot M\left(\boldsymbol{\theta}\right)}{M\left(X, \mathbf{d}\right)}, \tag{9}$$

where $M\left(X, \mathbf{d}|\boldsymbol{\theta}\right)\ \mathbf{d}|$ is independent of $\left(X, \mathbf{d}\right)\ \mathbf{d}$, $M\left(X, \mathbf{d}|\boldsymbol{\theta}\right)\ \mathbf{d}|$ can be considered a constant, allowing Eq. (9) to be simplified to:

$$\boldsymbol{\theta}^* = \arg\ \max\ M\left(X, \mathbf{d}|\boldsymbol{\theta}\right) \cdot M\left(\boldsymbol{\theta}\right). \tag{10}$$

By leveraging the monotonicity of the logarithmic function, Eq. (10) can be expressed as:

$$\boldsymbol{\theta}^* = \arg\ \max\left(\log\ M\left(\mathrm{X}, \mathbf{d}|\boldsymbol{\theta}\right) \cdot \log\ M\left(\boldsymbol{\theta}\right)\right), \tag{11}$$

Eq. (11) requires the construction of two probability density functions. $M\left(X, \mathbf{d}|\boldsymbol{\theta}\right)\ \mathbf{d}|$ represents the distance between the predicted distribution and the true distribution. The KL (Kullback–Leibler) divergence is commonly chosen to measure the distance between the two distributions. The aforementioned probability densities can be expressed as:

$$M\left(\mathrm{X}, \mathbf{d}|\boldsymbol{\theta}\right) = \sum_t \mathbf{d}_{\mathrm{t}} \ln \frac{\mathbf{d}_t}{\mathbf{d}_t^*}, \tag{12}$$

where $\mathbf{d}^*$ represents the predicted label. For the prior term, in neural networks, it follows a normal Gaussian distribution, and the probability density is expressed as:

$$M\left(\boldsymbol{\theta}\right) = \frac{1}{\sqrt{2\pi\delta^2}} Exp\left(-\frac{\theta^2}{2\delta^2}\right). \tag{13}$$

Therefore, the solution formula for parameter $\boldsymbol{\theta}^*$ can be further written as:

$$\boldsymbol{\theta}^* = \arg\ \min\left(\sum_t \mathbf{d}_t \ln \frac{\mathbf{d}_t}{\mathbf{d}_t^*} + \lambda \left\|\boldsymbol{\theta}\right\|_2^2\right). \tag{14}$$

In the loss function section, the Gaussian distribution loss is defined as:

$$KL = \sum_t \mathbf{d}_{\mathrm{t}} \ln \frac{\mathbf{d}_t}{\mathbf{d}_t^*} + \lambda \left\|\boldsymbol{\theta}\right\|_2^2, \tag{15}$$

where λ is the coefficient of L_2 regularization.KL represents the Kullback–Leibler Divergence.L_2 regularization is incorporated to prevent excessive growth of training parameters, thereby avoiding overfitting of the model. Then, construct the regression component, calculate the category based on the regression component, restore each category to the predicted angle, and compute the mean squared error loss between the predicted angle and the true angle. The mean squared error is calculated as follows:

$$MSE\left(y, \hat{y}\right) = \frac{1}{N} \sum_{i=0}^{N} \left(y_i - \hat{y}_i | x_i, \boldsymbol{\theta}\right)^2, \tag{16}$$

Finally, the two components are balanced using α and β to obtain the total loss $\mathcal{L}$:

$$\mathcal{L} = \alpha \cdot KL\left(\mathbf{d}, \hat{\mathbf{d}}\right) + \beta \cdot MSE\left(y, \hat{y}\right). \tag{17}$$

4. Experiments

4.1. Generally setting

4.1.1. Datasets

To thoroughly validate the proposed unconstrained appearance-based gaze estimation method, the network model undergoes experimental training and validation on the MPIIFaceGaze [33] and Gaze360

[34] gaze estimation datasets. The dataset was first introduced in [35] and named the MPIIGaze dataset, initially containing only eye images and not facial images. In subsequent research [33,36], the authors added facial images and landmark annotations to the original dataset to create the MPIIFaceGaze dataset. Based on prior work on MPIIFaceGaze [36,37], we employ a leave-one-person cross-validation strategy to evaluate the performance of each method. Additionally, we adhere to the standard evaluation split of the MPIIFaceGaze dataset and report the performance of each model. As the MPIIFaceGaze and Gaze360 datasets comprise a large number of images from varied backgrounds, illumination conditions, head positions, and gaze directions, this diversity ensures that the model can be thoroughly tested under various real-world conditions, enhancing the algorithm's generalizability.

- **MPIIFaceGaze** [33]: The MPIIGaze [35] dataset is the first to provide unconstrained data for gaze estimation in the wild. This dataset includes 15 subjects (9 males, 6 females, and 5 glasses wearers) in various daily laptop usage scenarios, where the target occasionally appears at random positions on the screen. The recorded data encompasses diverse environmental conditions, including indoor and outdoor settings, lighting variations, head poses, and positional changes, as well as overall recording quality. Given that the MPIIGaze [35] dataset provides cropped eye regions, this study uses its improved version, MPIIFaceGaze [33], which provides 3000 normalized full-face images for each subject.
- **Gaze360** [34]: A large-scale dataset for unconstrained 3D gaze estimation, Gaze360 uniquely incorporates a wide range of gaze and head poses, 3D gaze annotations, various indoor and outdoor settings, and diverse subjects categorized by age, gender, and race. The dataset comprises 172,000 images from 238 participants, with each image resolution being 3382 × 4096 pixels. The dataset was collected across 5 indoor locations (53 participants) and 2 outdoor locations (185 participants). The dataset comprises 58% female and 42% male participants. This dataset enables gaze estimation to push the boundaries of eye visibility, in some cases corresponding to a gaze yaw of ±140°.
- **EyeDiap** [38]: Participants sat facing a Kinect depth camera, tracking a randomly moving ping-pong ball with their eyes, which was manipulated by an experimenter. The entire process was recorded using the depth camera. Post-collection, eye and ball positions were annotated on RGB videos, mapped onto the 3D point cloud from the depth camera, and their difference yielded the gaze direction. Sessions for participants 14, 15, and 16 were recorded twice under varied conditions (A and B) including date, lighting, and distance from the camera. To assess robustness against head pose variations, participants were instructed to gaze at targets while keeping their head still (Static, S) or moving it (Dynamic, M). Sixteen participants took part, comprising 12 males and 4 females. Files include frame-wise head pose, eye and ball position data in both 2D and 3D, screen coordinates, and calibration parameters for RGB Kinect, depth, and RGB HD cameras.

4.1.2. Evaluation metrics

Angular error ϵ is utilized as a performance metric for the model, indicating the angular difference between the predicted and actual gaze directions. It intuitively quantifies the deviation between the predicted and actual directions. The smaller the angular error, the greater the accuracy of the algorithm. The specific method of calculation is as follows:

$$\epsilon = \frac{1}{n}\sum_{i}^{n} \arccos \frac{\langle \mathbf{g}_i, \hat{\mathbf{g}}_i \rangle}{|\mathbf{g}_i|\,|\hat{\mathbf{g}}_i|}, \tag{18}$$

where ϵ denotes the three-dimensional gaze angular error, $\mathbf{g}_i$ is the actual gaze direction, and $\hat{\mathbf{g}}_i$ is the model's predicted gaze direction. The three-dimensional gaze direction is derived from the two-dimensional gaze direction, and the method of calculation is as follows:

$$\mathbf{g}_i = \left(-\cos\boldsymbol{\alpha}_i \times \sin\boldsymbol{\beta}_i, -\sin\boldsymbol{\alpha}_i, -\cos\boldsymbol{\alpha}_i \times \cos\boldsymbol{\beta}_i\right), \tag{19}$$

$$\hat{\mathbf{g}}_i = \left(-\cos\hat{\boldsymbol{\alpha}}_i \times \sin\hat{\boldsymbol{\beta}}_i, -\sin\hat{\boldsymbol{\alpha}}_i, -\cos\hat{\boldsymbol{\alpha}}_i \times \cos\hat{\boldsymbol{\beta}}_i\right), \tag{20}$$

where $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ respectively denote the pitch and yaw angles of the actual two-dimensional gaze direction, while $\hat{\boldsymbol{\alpha}}_i$ and $\hat{\boldsymbol{\beta}}_i$ represent the pitch and yaw angles of the model's predicted two-dimensional gaze direction.

4.1.3. Training details

In this work, we employ ResNet-50 [39] as the backbone network of the model. We train the network using the Adam optimization algorithm [40], starting with an initial learning rate of 0.0001, which is further reduced by a factor of 10 every 10 epochs. Our network architecture is implemented using Pytorch [41] and runs on a workstation equipped with an Nvidia RTX 8000 GPU, which boasts 48 GB of VRAM. The batch sizes used for training on the MPIIFaceGaze and Gaze360 datasets are set at 32 and 16, respectively.

4.2. Effect of different backbone

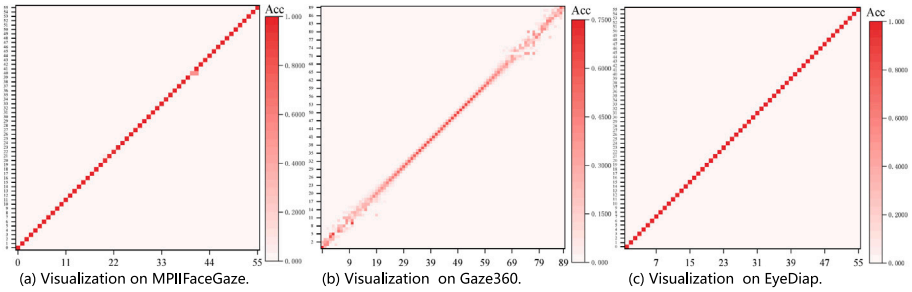
As shown in Table 1, ResNet18 [42], ResNet50 [42], ResNet152 [42], and Transformer [43] were chosen as the backbone networks for feature extraction. Training and testing were conducted on the MPIIFaceGaze and Gaze360 datasets. The results indicate that ResNet50 has strong feature extraction capabilities and performs excellently, whereas the Transformer model, as the backbone network, shows poor gaze estimation accuracy. Transformer models typically have a large number of parameters and complex structures, making it difficult for them to effectively learn and generalize from the relatively small datasets commonly used in gaze estimation scenarios. In contrast, convolutional neural networks like ResNet, with their local receptive fields and weight-sharing properties, can more effectively learn features and avoid overfitting. ResNet18, with its shallow network depth, has limited expressive power and feature extraction capabilities, resulting in poor performance in gaze estimation. Conversely, ResNet152, due to its excessive depth, has a large model capacity, making it prone to overfitting. Therefore, ResNet50, compared to ResNet18, has a deeper network extraction capability, enabling it to extract more complex features and enhance the model's expressive power, and it does not have excessive redundant parameters that increase the computational cost, as seen in ResNet152. Experimental results show that ResNet50 performs better as a backbone for feature extraction. Therefore, this study selects ResNet50 as the backbone for gaze feature extraction.

Table 1
Performance of the different backbone on the MPIIFaceGaze, Gaze360 and EyeDiap datasets.

Methods	MPIIFaceGaze	Gaze360	EyeDiap
ResNet18 [42]	5.06	12.79	6.59
ResNet152 [42]	4.87	11.85	6.36
Vit [43]	6.77	15.32	6.72
Swin-transformer [44]	4.61	12.35	5.77
ResNet50 [42]	**3.83**	**10.66**	**4.83**

4.3. Accuracy analysis of coarse classification

To further explore the efficiency of the internal classification mechanism of the Coarse Model, the training results on two public datasets were collected during the training process. For the MPIIFaceGaze dataset, the same strategy was adopted, and the results are shown in Fig. 4(a). It can also be found that overall, the accuracy of each class is relatively high, but class 41 is slightly lower. For the Gaze360



(a) Visualization on MPIIFaceGaze. (b) Visualization on Gaze360. (c) Visualization on EyeDiap.

Fig. 4. Experimental results of Coarse Classification on the different dataset.

Table 2
Performance of the ADGaze method on the MPIIFaceGaze, Gaze360 and EyeDiap datasets.

Methods	MPIIFaceGaze	Gaze360	EyeDiap
Mnist [35]	6.39	N/A	7.37
FullFace [33]	4.93	14.99	6.53
RT-Gene [45]	4.66	12.26	6.02
RCNN [46]	4.1	11.23	5.43
Dilated-Net [47]	4.42	13.73	6.19
Gaze360 [34]	4.06	11.04	5.36
GazeNet [36]	5.76	N/A	6.79
CA-Net [48]	4.27	11.20	5.27
GazeTR-Pure [49]	4.74	13.58	5.72
GTiT-Hybrid [50]	4.11	11.20	5.35
SAZE [24]	3.89	N/A	**4.42**
ADGaze_S	3.83	10.66	4.83
ADGaze	**3.62**	**10.63**	4.67

dataset, with its 90 classification categories, the data under the Pitch angle was also collected, as shown in Fig. 4(b). It can be seen that although the overall accuracy is lower than the previous two datasets, the misclassified categories are distributed around the correct categories, which indicates that ADGaze learns the differences in different angle directions through anisotropic label distribution and can also fully utilize surrounding instance information to improve the model's prediction ability, making the predicted labels as close to the true labels as possible.

4.4. Performance comparison

The primary aim of this section is to validate the effectiveness of the ADGaze model in gaze estimation tasks by conducting a comprehensive comparison with the eight latest gaze estimation methods. Including Euler angle regression methods(FullFace [33], RT-Gene [45], Gaze360 [34]), RCNN [46], extra information-utilized methods (Mnist [35], GazeNet [36], Dilated-Net [47], CA-Net [48], GazeTR-Pure [49], GTiT-Hybrid [50], SAZE [24]), and alternative ways to use Transformer (GazeTR [49]). This comparison encompasses not only direct benchmarks of performance indicators such as accuracy and robustness but also assesses the model's capability to handle various data types and complex scenarios. Through this series of comparative analyses, the goal is to showcase the advantages of the ADGaze model within the field of gaze estimation. These advantages likely include higher prediction accuracy, enhanced generalization capabilities, and robustness to various challenging conditions such as occlusions, extreme poses, and illumination changes. By comparing with some advanced gaze estimation methods on two publicly available gaze estimation datasets, the results demonstrate that the adaptive feature disentanglement gaze estimation network proposed in this work exhibits strong performance.

We use two commonly utilized public datasets, attempting both the statistical label distribution method (ADGaze_S) and the feature similarity-based label distribution method (ADGaze). These methods are compared against other approaches. Overall, this method demonstrates good performance on some datasets. Specifically, as shown in Table 2, it achieves an average angular error of 3.62° on the MPIIFaceGaze dataset and an average angular error of 10.63° on the Gaze360 dataset. Overall, the ADGaze_S model shows improvement over the backbone model on public datasets. Specifically, on the MPIIFaceGaze dataset, the ADGaze model has the smallest error, at 3.62°, making it the best-performing model among all.

It should be noted that the performance comparison of different models is influenced by many factors, such as the size and diversity of the dataset, the complexity of the model, and the parameter settings. Therefore, in practical applications, it is necessary to choose the most suitable model and parameter settings based on the specific situation. In this experiment, the ADGaze model performed well on both the MPIIFaceGaze and Gaze360 datasets, making it an effective method for eye gaze prediction.

4.5. Effect of Anisotropism Soft Label Construct (ASLC) model

To verify the effectiveness of the ASLC model, an ablation experiment was conducted. As shown in Table 3, after removing the ASLC Model, the ASLC model achieved an error of 3.92° on the MPIIFaceGaze dataset, an error of 10.72° on the Gaze360 dataset and an error of 5.47° on the EyeDiap dataset. When adding the statistical label distribution to the base model, the ADGaze_S model achieved an error of 3.83° on the MPIIFaceGaze dataset, an error of 10.66° on the Gaze360 dataset and an error of 4.83° on the EyeDiap dataset. We found that even if the statistical data (e.g., quantity) of two categories are the same, their distribution in the feature space can be very different. Feature similarity can account for the shape of this distribution, whereas statistical methods cannot. Therefore, the ADGaze model, which uses feature similarity-based label distribution, performs better, achieving an error of 3.62° on the MPIIFaceGaze dataset, an error of 10.63° on the Gaze360 dataset and an error of 4.67° on the EyeDiap dataset. Compared to the base model, both ADGaze_S and ADGaze achieved optimal results. This further indicates that directly fitting gaze angles make it difficult for the model to learn an effective feature representation. However, using the coarse-to-fine feature extraction framework proposed in this paper can effectively reduce the search efficiency and computational complexity of the solution space, obtaining a small angular search space through classification and then achieving more accurate gaze estimation through regression. This improves the model's accuracy on one hand while also reducing the model's computational complexity on the other (see Table 3).

4.6. Loss function discussion

To evaluate which loss functions can more effectively capture the relationships between neighboring samples when dealing with highly similar ocular images, thereby improving the accuracy of gaze estimation, we conducted comparative experiments with other commonly

Table 3
Ablation experiment of different components on MPIIFaceGaze, Gaze360 and EyeDiap datasets.

Methods	MPIIFaceGaze	Gaze360	EyeDiap
Baseline	5.24	13.95	5.79
w/o_{ASLC}	3.92	10.72	5.47
ADGaze_S	3.83	10.66	4.83
ADGaze	**3.62**	**10.63**	**4.67**

Table 4
Ablation experiment of different loss functions on MPIIFaceGaze, Gaze360 and EyeDiap datasets.

Loss function	MPIIFaceGaze	Gaze360	EyeDiap
$Cross-Entropy$	4.27	11.08	5.23
JS	4.13	10.91	5.27
KL	**3.62**	**10.63**	**4.67**

used loss functions, such as Jensen–Shannon divergence (JS divergence) and Cross-Entropy loss. Our experimental results indicated that the Kullback–Leibler divergence (KL divergence) yielded superior outcomes. As shown in Table 4, these findings highlight the effectiveness of KL divergence in enhancing the precision of gaze estimation under conditions of high ocular image similarity. This investigation underscores the importance of selecting an appropriate loss function to refine model training, especially in tasks requiring nuanced differentiation between closely related data points. The presented results in Table 4 provide a detailed comparison, offering insights into the advantages of utilizing KL divergence over alternative loss functions for this specific application.

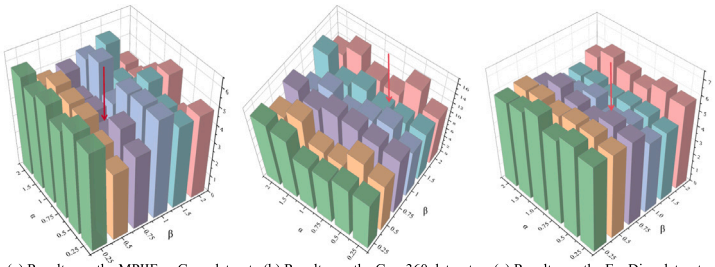
4.7. Parameter discussion

In the ADGaze model, the loss is obtained by weighting the total loss, which supervises the model's training. Thus, the choice of parameters will affect the training results. When selecting experimental schemes, the general dataset training and testing division standards for gaze estimation were adopted, and multiple batches of experiments with different parameters were conducted on two public datasets. As shown in Fig. 5, it can be observed that on the MPIIFaceGaze dataset, when the total loss weights are $\alpha = 1.5$ and $\beta = 0.5$, on the Gaze360 dataset, when the total loss weights are $\alpha = 1$ and $\beta = 0.75$, the ADGaze_S model achieved optimal performance, and on the EyeDiap dataset, when the total loss weights are $\alpha = 0.75$ and $\beta = 1.0$, the ADGaze_S model achieved optimal performance. As shown in Fig. 6, the performance of ADGaze with different α and β values after the same number of iterations is illustrated on two public datasets. Fig. 6(a) shows that when $\alpha = 1$ and $\beta = 1$, the CF model performs best on the MPIIFaceGaze dataset; Fig. 6(b) shows that when $\alpha = 0.8$ and $\beta = 0.6$, the ADGaze model performs best on the Gaze360 dataset. Experiments show that under the default parameters $\alpha = 1$ and $\beta = 1$; and Fig. 6(c) shows that when $\alpha = 0.75$ and $\beta = 1.5$, the ADGaze model performs best on the EyeDiap dataset. Experiments show that under the default parameters $\alpha = 1$ and $\beta = 1$, adjusting the values of the hyperparameters α and β can improve the model's performance. This is consistent with the analysis of the datasets in this paper, as the data distribution, while generally consistent, varies due to differences in the quantity, collection methods, and quality of different datasets. However, the distribution parameters are not necessarily identical, which is why the hyperparameters of the composite loss are not necessarily the same.

4.8. Visualizations

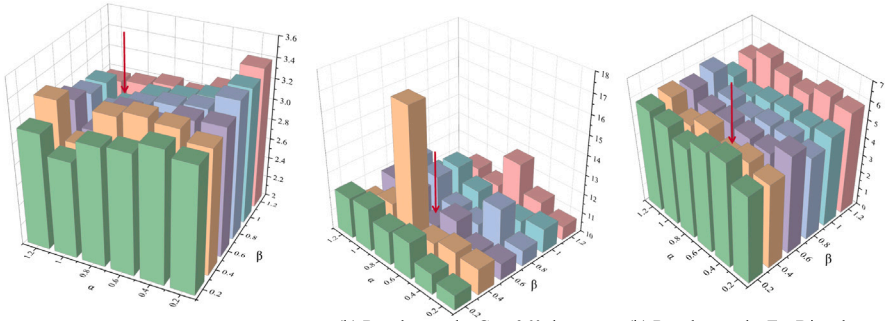
To further test the effectiveness of the model, this section visualizes the gaze directions estimated by the ADGaze_S and ADGaze models, as shown in Fig. 7(a). The experimental results show that the ADGaze_S model can provide predicted angles very close to the true values when processing both normal images and images affected by various distortions. Especially when dealing with complex situations such as occlusions, extreme angles, and lighting changes, the ADGaze_S model still maintains high accuracy and strong robustness, which is particularly important in gaze estimation tasks, as real-world application scenarios are often full of uncertainties and challenges. The ADGaze results are shown in Fig. 8, where the red lines indicate the predicted direction and the green lines indicate the true gaze direction. Fig. 7(b)(1), (2), and (3) are images selected from the MPIIFaceGaze [21] dataset, where it can be seen that the sample images are greatly affected by lighting, especially Fig. 7(b)(4), which is in a very dark environment. However, during the actual prediction process, it was found that the ADGaze model showed relatively robust results in such harsh environments. Fig. 7(b)(4), (5), and (6) are images selected from the Gaze360 [20] test set, where the image clarity is low. Fig. 7(b)(4) shows very low image clarity, Fig. 7(b)(5) shows an extreme gaze direction, and Fig. 7(b)(6) shows not only a blurry image but also a more challenging gaze direction compared to Fig. 7(b)(5). It is worth mentioning that in Fig. 7(b)(4), (5), and (6), although the eye features are severely affected by the pixel and angle of the test camera, the final gaze estimation results are still relatively close to the true values. Even when the test set images are blurry, severely affected by lighting, and have challenging gaze angles, this study captures neighborhood information by utilizing gaze anisotropy, thereby enhancing the robustness of gaze direction prediction.

To further explore the reasons for the robustness of the feature extraction by the model, this paper uses Grad-CAM [51] to visualize the entire face image. It can be observed that for the same person, the corresponding areas of the heatmap are concentrated near the eye region. When the angles are different, the activation areas may vary; when facing forward, the activation areas are in the double-eye region, and when sideways, the activation areas are in the single-eye region. This indicates that the model extracts effective regional features during training, ignoring irrelevant areas such as image contours, thereby enhancing the model's performance and reducing errors. As shown in Figs. 8 and 9, the features of the ADGaze_S and ADGaze models are mainly concentrated on one or both eyes. Specifically, as shown in Fig. 9(b), some sample heatmaps of the ADGaze model on the Gaze360 dataset show that in Fig. 9(b)(4), the model focuses on the double-eye region but extracts features from the right eye more strongly. In Fig. 9(b), the gaze is skewed, with significant feature extraction from the left eye and nose region. In Fig. 9(b) (6), with a large head pose variation, features from both the nose and left eye regions are extracted. As shown in Figs. 8(a) and 9(a), ADGaze_S focuses more on the single-eye region, whereas ADGaze focuses on both the single-eye region and other areas. Therefore, the ADGaze model can capture more neighborhood information. The heatmap effect of some samples of the ADGaze model on the MPIIFaceGaze dataset is shown in Fig. 9(a)(4) and (b)(5), where the samples are wearing glasses. However, the features extracted by the ADGaze model vary with the gaze direction; in Fig. 9(a)(6), it focuses only on the right eye region, while in Fig. 9(a)(5), it focuses on the double-eye region but with greater attention to the right eye. Fig. 9(a)(6) also shows feature extraction from both eyes, along with information from the cheek and head pose-related areas. Although the model is designed to handle highly similar eye images and, in most cases, the feature regions are concentrated around the eyes, under extreme head poses, such as in Figs. 8(c)(5) and 9(c)(5), even slight changes can lead to errors in the classification or regression stages. This results in the heatmap regions not focusing on the eye areas but rather on the nose area. Although the model encounters challenges with feature extraction region shifts under extreme head poses, our future work will focus on augmenting the dataset with more samples of extreme head poses, either by collecting real-world examples or generating synthetic ones using generative models. From the above figures,



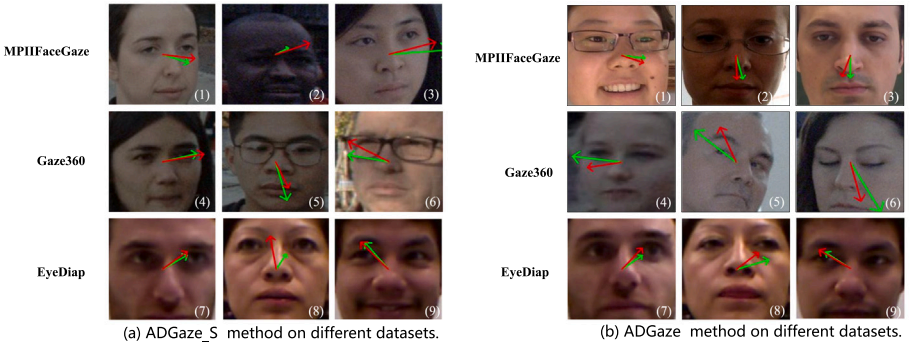
(a) Results on the MPIIFaceGaze dataset. (b) Results on the Gaze360 dataset. (c) Results on the EyeDiap dataset.

Fig. 5. Experimental results of ADGaze_S under varying α and β settings on the different dataset.



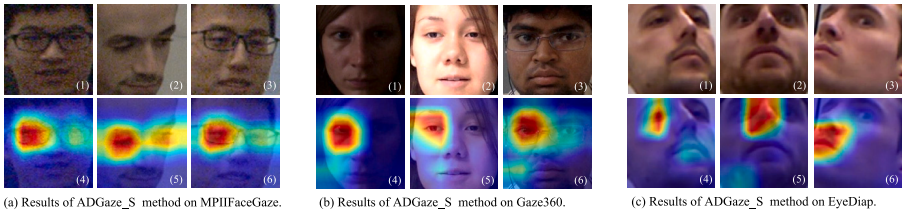
(a) Results on the MPIIFaceGaze dataset. (b) Results on the Gaze360 dataset. (b) Results on the EyeDiap dataset.

Fig. 6. Experimental results of ADGaze under varying α and β settings on the different dataset.



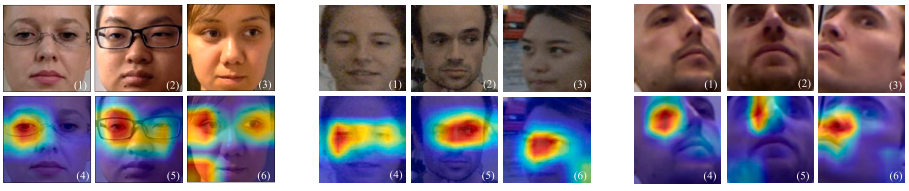
(a) ADGaze_S method on different datasets. (b) ADGaze method on different datasets.

Fig. 7. Visualization results of methods on different datasets.



(a) Results of ADGaze_S method on MPIIFaceGaze. (b) Results of ADGaze_S method on Gaze360. (c) Results of ADGaze_S method on EyeDiap.

Fig. 8. Grad-CAM results generated by ADGaze_S method for different samples.



(a) Results of ADGaze method on MPIIFaceGaze. (b) Results of ADGaze method on Gaze360. (c) Results of ADGaze method on EyeDiap.

Fig. 9. Grad-CAM results generated by ADGaze method for different samples.

it can be seen that even when there are issues such as blurriness and lighting in the data, the proposed method can utilize domain feature information, so the extracted gaze features are concentrated in the eye region, further enhancing the model's robustness. From the above figures, it can be seen that even when there are issues such as blurriness and lighting in the data, the proposed method can utilize domain feature information, so the extracted gaze features are concentrated in the eye region, further enhancing the model's robustness.

5. Conclusion

This study introduces ADGaze, an anisotropic Gaussian label distribution learning framework to address the problem of gaze estimation. By analyzing eye images and utilizing the cosine similarity function, ADGaze reveals the anisotropic relationship between gaze directions. The framework's coarse-to-fine regression approach effectively exploits this characteristic by first predicting the coarse range of gaze direction and then fine-tuning the estimation within this range to improve prediction accuracy. Our experiments on two public datasets demonstrate that ADGaze outperforms state-of-the-art methods in terms of accuracy. However, one limitation of this work is its focus solely on gaze direction while overlooking variations in head pose, which may restrict the model's ability to generalize to more complex head poses. Despite this limitation, the proposed approach offers a novel perspective on gaze estimation by emphasizing the anisotropic nature of eye movements. Future research will aim to address this issue by developing a more comprehensive anisotropic distribution model that incorporates both gaze direction and head pose. This improvement is expected to broaden ADGaze's applicability to more diverse scenarios. Additionally, the findings have potential applications in areas such as attention tracking and intelligent human–computer interaction, providing valuable insights for designing gaze estimation models in real-world applications.

CRediT authorship contribution statement

Duantengchuan Li: Writing – review & editing, Writing – original draft, Software, Project administration, Methodology, Formal analysis. **Shutong Wang:** Writing – review & editing, Writing – original draft, Software, Conceptualization. **Wanli Zhao:** Writing – original draft, Software, Data curation, Conceptualization, Writing – review & editing. **Lingyun Kang:** Writing – review & editing, Supervision, Methodology. **Liangshan Dong:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Jiazhang Wang:** Writing – review & editing, Visualization, Methodology, Conceptualization. **Xiaoguang Wang:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization.

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

The authors sincerely thank anonymous reviewers for their constructive comments, which helped improve this article. This work was supported by the National Key R&D Program of China (No. 2022YFF0901902), the National Natural Science Foundation of China (No. 62307034), the Hubei Provincial Natural Science Foundation, China (No. 2023AFB359), the Research Funds from the Social Science Foundation of Jiangxi Province (No. 24JY14) and 2024 Jiangxi University Humanities and Social Sciences Research Youth Program (No. JY24203).

Data availability

Data will be made available on request.

References

[1] Y. Sun, J. Zeng, S. Shan, Gaze estimation with semi-supervised eye landmark detection as an auxiliary task, Pattern Recognit. 146 (2024) 109980.
[2] R. Valenti, N. Sebe, T. Gevers, Combining head pose and eye location information for gaze estimation, IEEE Trans. Image Process. 21 (2) (2012) 802–815.
[3] Q. Ji, X. Yang, Real-time eye, gaze, and face pose tracking for monitoring driver vigilance, Real-Time Imaging 8 (5) (2002) 357–377.
[4] M.X. Huang, T.C. Kwok, G. Ngai, H.V. Leong, S.C. Chan, Building a self-learning eye gaze model from user interaction data, in: Proceedings of the 22nd ACM International Conference on Multimedia, 2014, pp. 1017–1020.
[5] Y. Zhang, A. Bulling, H. Gellersen, Pupil-canthi-ratio: a calibration-free method for tracking horizontal gaze direction, 2014.
[6] Y. Cheng, F. Lu, Gaze estimation using transformer, in: 2022 26th International Conference on Pattern Recognition, ICPR, 2022, pp. 3341–3347.
[7] V. Nagpure, K. Okuma, Searching efficient neural architecture with multi-resolution fusion transformer for appearance-based gaze estimation, in: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, 2023, pp. 890–899.
[8] H. Liu, S. Fang, Z. Zhang, D. Li, K. Lin, J. Wang, Mfdnet: Collaborative poses perception and matrix Fisher distribution for head pose estimation, IEEE Trans. Multimed. 24 (2022) 2449–2460.
[9] W. Zhao, S. Wang, X. Wang, D. Li, J. Wang, C. Lai, X. Li, Dadl: Double asymmetric distribution learning for head pose estimation in wisdom museum, J. King Saud Univ. - Comput. Inf. Sci. 36 (1) (2024) 101869.
[10] B. Liu, D. Li, J. Wang, Z. Wang, B. Li, C. Zeng, Integrating user short-term intentions and long-term preferences in heterogeneous hypergraph networks for sequential recommendation, Inf. Process. Manage. 61 (3) (2024) 103680.
[11] H. Liu, C. Zheng, D. Li, X. Shen, K. Lin, J. Wang, Z. Zhang, Z. Zhang, N.N. Xiong, Edmf: Efficient deep matrix factorization with review feature learning for industrial recommender system, IEEE Trans. Ind. Informatics 18 (7) (2022) 4361–4371.
[12] D. Li, Y. Gao, Z. Wang, H. Qiu, P. Liu, Z. Xiong, Z. Zhang, Homogeneous graph neural networks for third-party library recommendation, Inf. Process. Manage. 61 (6) (2024) 103831, http://dx.doi.org/10.1016/j.ipm.2024.103831.
[13] F. Shi, D. Li, X. Wang, B. Li, X. Wu, Tgformer: A graph transformer framework for knowledge graph embedding, IEEE Trans. Knowl. Data Eng. 37 (1) (2025) 526–541, http://dx.doi.org/10.1109/TKDE.2024.3486747.
[14] D. Li, T. Xia, J. Wang, F. Shi, Q. Zhang, B. Li, Y. Xiong, Sdformer: A shallow-to-deep feature interaction for knowledge graph embedding, Knowl.-Based Syst. 284 (2024) 111253.
[15] K. Lin, Y. Li, S. Chen, D. Li, X. Wu, Motion planner with fixed-horizon constrained reinforcement learning for complex autonomous driving scenarios, IEEE Trans. Intell. Veh. 9 (1) (2024) 1577–1588, http://dx.doi.org/10.1109/TIV.2023.3273857.
[16] K. Wang, R. Zhao, H. Su, Q. Ji, Generalizing eye tracking with Bayesian adversarial learning, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11899–11908.
[17] K.A. Funes Mora, J.-M. Odobez, Geometric generative gaze estimation (G3E) for remote RGB-D cameras, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1773–1780.
[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 6000–6010.
[19] H. Cheng, Y. Liu, W. Fu, Y. Ji, L. Yang, Y. Zhao, J. Yang, Gazing point dependent eye gaze estimation, Pattern Recognit. 71 (2017) 36–44.
[20] X. Geng, Label distribution learning, IEEE Trans. Knowl. Data Eng. 28 (7) (2016) 1734–1748.
[21] K. Smith-Miles, X. Geng, Revisiting facial age estimation with new insights from instance space analysis, IEEE Trans. Pattern Anal. Mach. Intell. 44 (5) (2022) 2689–2697.
[22] X. Geng, X. Qian, Z. Huo, Y. Zhang, Head pose estimation based on multivariate label distribution, IEEE Trans. Pattern Anal. Mach. Intell. 44 (4) (2022) 1974–1991.
[23] K. Wang, Q. Ji, 3D gaze estimation without explicit personal calibration, Pattern Recognit. 79 (2018) 216–227.
[24] S. Kim, W.-J. Nam, S.-W. Lee, Appearance debiased gaze estimation via stochastic subject-wise adversarial learning, Pattern Recognit. 152 (2024) 110441.
[25] C. Gou, Y. Yu, Z. Guo, C. Xiong, M. Cai, Cascaded learning with transformer for simultaneous eye landmark, eye state and gaze estimation, Pattern Recognit. 156 (2024) 110760.
[26] K.R. Park, J.J. Lee, J. Kim, Gaze position detection by computing the three dimensional facial positions and motions, Pattern Recognit. 35 (11) (2002) 2559–2569.

[27] M.M. Marefat, L. Wu, C.C. Yang, Gaze stabilization in active vision—I. Vergence error extraction, Pattern Recognit. 30 (11) (1997) 1829–1842.

[28] Y. Yu, G. Liu, J.-M. Odobez, Deep multitask gaze estimation with a constrained landmark-gaze model, in: Computer Vision – ECCV 2018 Workshops, 2019, pp. 456–474.

[29] X. Zhao, Y. An, L. Qi, X. Geng, Scalable label distribution learning for multi-label classification, IEEE Trans. Neural Netw. Learn. Syst. (2024) 1–15.

[30] X. Zhao, Y. An, N. Xu, X. Geng, Variational continuous label distribution learning for multi-label text classification, IEEE Trans. Knowl. Data Eng. 36 (6) (2024) 2716–2729.

[31] H. Xu, X. Liu, Q. Zhao, Y. Ma, C. Yan, F. Dai, Gaussian label distribution learning for spherical image object detection, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 1033–1042.

[32] X. Li, G. Luo, W. Wang, K. Wang, S. Li, Curriculum label distribution learning for imbalanced medical image segmentation, Med. Image Anal. 89 (2023) 102911.

[33] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, It's written all over your face: Full-face appearance-based gaze estimation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 2299–2308.

[34] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, A. Torralba, Gaze360: Physically unconstrained gaze estimation in the wild, in: 2019 IEEE/CVF International Conference on Computer Vision, 2019, pp. 6911–6920.

[35] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, Appearance-based gaze estimation in the wild, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4511–4520.

[36] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, MPIIGaze: Real-world dataset and deep appearance-based gaze estimation, IEEE Trans. Pattern Anal. Mach. Intell. 41 (1) (2019) 162–175.

[37] Y. Cheng, X. Zhang, F. Lu, Y. Sato, Gaze estimation by exploring two-eye asymmetry, IEEE Trans. Image Process. 29 (2020) 5259–5272.

[38] K.A. Funes Mora, F. Monay, J.-M. Odobez, EYEDIAP: A Database for the Development and Evaluation of Gaze Estimation Algorithms from RGB and RGB-D Cameras, Association for Computing Machinery, New York, NY, USA, 2014.

[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[40] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, ICLR, San Diega, CA, USA, 2015.

[41] A. Paszke, S. Gross, Massa, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc., 2019.

[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778.

[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.

[44] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 10012–10022.

[45] T. Fischer, H.J. Chang, Y. Demiris, RT-GENE: Real-time eye gaze estimation in natural environments, in: Computer Vision – ECCV 2018, 2018, pp. 339–357.

[46] C. Palmero, J. Selva, M.A. Bagheri, S. Escalera, Recurrent CNN for 3D gaze estimation using appearance and shape cues, in: British Machine Vision Conference 2018, 2018, p. 251.

[47] Z. Chen, B.E. Shi, Appearance-based gaze estimation using dilated-convolutions, in: C. Jawahar, H. Li, G. Mori, K. Schindler (Eds.), Computer Vision – ACCV 2018, 2019, pp. 309–324.

[48] Y. Cheng, S. Huang, F. Wang, C. Qian, F. Lu, A coarse-to-fine adaptive network for appearance-based gaze estimation, Proc. AAAI Conf. Artif. Intell. 34 (07) (2020) 10623–10630.

[49] Y. Cheng, F. Lu, Gaze estimation using transformer, in: 2022 26th International Conference on Pattern Recognition, 2022, pp. 3341–3347.

[50] C. Wu, H. Hu, K. Lin, Q. Wang, T. Liu, G. Chen, Attention-guided and fine-grained feature extraction from face images for gaze estimation, Eng. Appl. Artif. Intell. 126 (2023) 106994.

[51] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision, 2017, pp. 618–626.

Duantengchuan Li received the M.S. degree from Central China Normal University, Wuhan, China, in 2021, and the Ph.D. degree in software engineering with the School of Computer Science, Wuhan University, Wuhan, China, in 2024. He is currently an Associate Researcher with the School of Information Management, Wuhan University, Wuhan, China. His research interests include computer vision, recommender system, representation learning, knowledge graph, intention recognition etc.

Shutong Wang received the M.S. degree from Central China Normal University, Wuhan, China, in 2024. He is currently working toward the Ph.D. degree in Management Science and Engineering at the School of Information Management, Wuhan University. His research interests include computer vision, human–computer interaction.

Wanli Zhao received his M.S. degree from Central China Normal University, Wuhan, China, in 2023. Currently, he is pursuing his Ph.D. at the School of Information Management, Wuhan University, Wuhan, China. Zhao's research interests primarily focus on computer vision and representation learning. In the realm of computer vision, Zhao is exploring advanced techniques for image and video analysis, aiming to develop more robust algorithms for object recognition and motion tracking.

Lingyun Kang received her Ph.D. from Central China Normal University, Wuhan, China, in 2023. Currently, she is an instructor at the College of Journalism and Communication at Jiangxi Normal University, Nanchang 330022, China. Concurrently, she is engaged in postdoctoral research at the School of Psychology within the same university. Her research interests lie in the fields of learning analytics and collaborative learning.

Liangshan Dong received the Ph.D. degree from Central China Normal University, Wuhan, China, in 2021. He is currently working with the School of Physical Education at China University of Geosciences (Wuhan). His research interests encompass a broad spectrum of topics, including pattern recognition, exercise intervention, and computer vision.

Jiazhang Wang received the Ph.D. degree in computational photography from the ECE Department, Northwestern University, Evanston, IL, USA. His research interests include pattern recognition, computer vision, and representation learning.

Xiaoguang Wang received the M.S. degree in informatics from Wuhan University in 2004, and the Ph.D. degree in Management Science and Engineering from the same university in 2007. Currently, he serves as a professor in the School of Information Management at Wuhan University. His research interests are diverse and encompass areas such as pattern recognition, computer vision, knowledge organization, semantic publishing, and digital humanities.