

Gaze Label Alignment: Alleviating Domain Shift for Gaze Estimation

Guanzhong Zeng*, Jingjing Wang*, Zefu Xu, Pengwei Yin, Wenqi Ren[†], Di Xie, Jiang Zhu

Hikvision Research Institute, Hangzhou, China

{zengguanzhong, wangjingjing9, xuzezu, yinpengwei, renwenqi, xiedi, zhujiang.hri}@hikvision.com

Abstract

Gaze estimation methods encounter significant performance deterioration when being evaluated across different domains, because of the domain gap between the testing and training data. Existing methods try to solve this issue by reducing the deviation of data distribution, however, they ignore the existence of label deviation in the data due to the acquisition mechanism of the gaze label and the individual physiological differences. In this paper, we first point out that the influence brought by the label deviation cannot be ignored, and propose a gaze label alignment algorithm (GLA) to eliminate the label distribution deviation. Specifically, we first train the feature extractor on all domains to get domain invariant features, and then select an anchor domain to train the gaze regressor. We predict the gaze label on remaining domains and use a mapping function to align the labels. Finally, these aligned labels can be used to train gaze estimation models. Therefore, our method can be combined with any existing method. Experimental results show that our GLA method can effectively alleviate the label distribution shift, and SOTA gaze estimation methods can be further improved obviously.

Introduction

The human gaze provides a wealth of information, highly accurate gaze estimation can strongly support many applications, such as human-computer interaction (Liu et al. 2021a), augmented reality (Murphy-Chutorian and Trivedi 2010), driver monitoring systems (Shah et al. 2022) and saliency prediction (Chang et al. 2019). Recently, appearance-based gaze estimation methods (Liu et al. 2019; Wang et al. 2022b) have achieved remarkable results with the development of deep learning. However, these methods obtain promising performance in within-domain evaluations but suffer from dramatic degradation in cross-domain evaluations due to the domain gap (Yin et al. 2024c).

To improve the performance in cross-domain scenarios, previous methods leverage domain adaptation approaches to reduce the domain gap. However, these methods (Bao et al. 2022; Wang et al. 2022b) require target domain samples for jointly optimizing the network with source domain data,

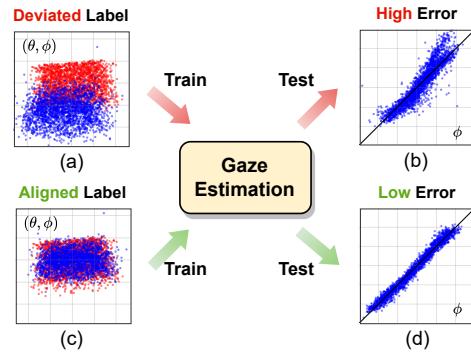


Figure 1: The red scatters in (a) and (c) mean the labels of source domain A and the corresponding blue scatters are the labels of source domain B . (b) and (d) are the scatter plots of estimated label (X-axis) and ground truth label (Y-axis).

which limits its application in real world, since the source data is usually not accessible due to data privacy, and it is not feasible to collect large amount of target domain data. To eliminate the dependency on target domain data during training, recently, some domain generalization approaches have been proposed. These methods (Cheng and Bao 2022; Yin et al. 2024c) do not need the target domain data, and learn domain invariant features by removing the gaze-irrelevant domain features.

Previous methods (Li et al. 2018, 2020; Wang et al. 2022a) only align the data distribution among domains and neglect the label distribution deviation. As it is infeasible to label the gaze direction manually, currently most methods get the gaze directions by building a gaze label acquisition system. The different internal and external parameters of different label acquisition systems lead to label shifts among different datasets. Moreover, even in the same label acquisition system, among different individuals, there also exists individual deviation due to the differences in eye shapes and inner eye structures. The deviation of the label causes the learned regressor to be divergent and have a relatively high error as shown in Fig. 1. In this paper, we point out the label distribution shift is also one of the main reasons causing performance degradation which has not been found by previous researchers in gaze estimation.

*These authors contributed equally.

[†]Corresponding author.

To align the label distributions, we propose a gaze label alignment algorithm. Specifically, we firstly train the feature extractor on all domains to get domain invariant features. Secondly, we select one of the domains as the anchor domain, and train the gaze regressor on the anchor domain. Thirdly, we predict the gaze label on remaining domains, then use the predicted labels and the ground truth labels on each domain to learn a mapping function to align their labels with the anchor domain. These two steps can be repeated several times. Finally, the data of all domains and aligned labels can be used to train the final gaze estimation model.

The main contributions can be summarized as follows:

- We are the first to point out that the label distribution shift also affects the domain generalization error, and should be aligned for better performance.
- We propose a gaze label alignment method (GLA) to align the label distribution shift among domains, which can be combined with any gaze estimation method.
- Experimental results demonstrate that our GLA method can effectively alleviate the label distribution shift. By aligning the label distribution, all the SOTA gaze estimation methods can be further improved obviously.

Related Works

Appearance-based Gaze Estimation

Appearance-based gaze estimation models directly regress gaze direction from face images. Since gaze labels cannot be manually annotated and data collection is expensive, in order to improve the performance of gaze estimation, early works prefer to learn more robust gaze features by relying on human prior knowledge. Then, Cheng et al. (Cheng et al. 2020) designed a coarse-to-fine strategy to refine the gaze features, and Biswas et al. (Biswas et al. 2021) assigned different attention weights to left and right eye features. Although these methods have achieved good within-domain performance on limited training data, they have high generalization errors on cross-domains (Xu, Wang, and Lu 2023).

Gaze Generalization and Adaptation

Due to the complex changes in head pose, illumination and face appearance in the real world. There are obvious domain gap between source domain and target domain. Previous methods focused on eliminating the data distribution shift and obtaining domain-invariant gaze features. For example, Xu et al. (Xu, Wang, and Lu 2023) generated noisy data through adversarial attacks, endowing the model generalization ability. Yin et al. (Yin et al. 2024a) utilized NeRF to generate diverse data to improve the performance in target domains. Recently, Yin et al. (Yin et al. 2024c,b) leveraged pre-trained vision-language models to introduce rich general knowledge and handle diverse target domains. However, they overlook label distribution shift between gaze domains, which also weakens the generalization capabilities of gaze estimation models. Therefore, we analyze the root causes and impacts of label domain shift and propose a label alignment method to reduce the difference in label distribution.

Method

Preliminaries

A gaze estimation network consists of two parts: a feature extractor F which outputs the gaze features $X = F(I)$ based on input image I , and a gaze regressor W which estimates the gaze direction $Y = W(X)$ based on input features X . Nevertheless, even the gaze estimation model achieves competitive performance in within-domain evaluation, the test error still has a high inter person variance in cross-domain evaluation (Zhang, Yao, and Cai 2018). This is caused by many factors including dependencies on head poses, large eye shape variabilities, and only very subtle eye appearance changes when looking at targets separated by such small gaze angle differences (Liu et al. 2019). Thus, previous domain generalization methods try to find a robust feature extractor F , which can extract invariant features against data distribution shift among domains. However, they neglect there is label deviation in gaze estimation due to the gaze label acquisition systems and individual physiological differences in the eyeball. This may affect the learning of the gaze regressor and hamper the final generalization performance.

More formally, let \mathcal{X} be the feature space and \mathcal{Y} the label space, a domain is defined as a joint distribution $P(X, Y)$ on $\mathcal{X} \times \mathcal{Y}$. Then $P(X, Y)$ can be decomposed into $P(X, Y) = P(Y|X)P(X)$, where $P(X)$ is the marginal distribution on X , and $P(Y|X)$ is the posterior distribution of Y given X . Previous domain generalization methods try to find a transformation of the data to extract feature X , which minimizes the difference between marginal distributions $P(X)$ of domains, while neglect the difference between $P(Y|X)$ which is caused by the gaze label acquisition systems and individual physiological differences. We call $P(X)$ the data distribution, and $P(Y|X)$ the label distribution.

What Causes the Gaze Label Deviation?

Individual Physiological Differences The main reason for the gaze label deviation between different peoples is that the visual axis is not aligned with the optical axis (related to the observed iris) (Guestrin and Eizenman 2006). And such alignment differences are person-specific, with a standard deviation of 2 to 3 degrees among the population without eye problems (Liu et al. 2019). Said differently, in theory, images of two eyes with the same appearance but with different internal eyeball structure can correspond to different gaze directions, which will result in prediction uncertainties of gaze estimation model.

Measurement Deviation of Data Acquisition System In addition, since it is infeasible to label the gaze direction manually, currently most methods get the gaze directions by building a gaze label acquisition system. The measurement deviations of these label acquisition systems lead to label deviations between multiple gaze datasets.

Usually, the system consists of a screen and a camera, which is shown in Fig. 2. The screen is used to show a gaze target $t_s = (u, v)$ in the screen coordinate system (SCS), and a user is required to look at this gaze target. Meanwhile, a face image is captured by the camera. The gaze direction g

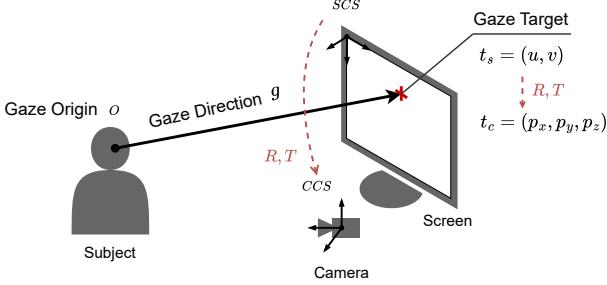


Figure 2: The schematic diagram of gaze direction. Gaze direction is originated from the gaze origin \mathbf{o} and intersects with the screen at gaze target \mathbf{t}_c .

can be calculated as follows, and treated as the ground truth gaze label in this image.

First, the gaze target $\mathbf{t}_s = (u, v)$ is transformed from the SCS to $\mathbf{t}_c(x_t, y_t, z_t)$ in the camera coordinate system (CCS) through the extrinsic parameters between the screen and camera, i.e. $\mathbf{t}_c = \mathbf{R}[u, v, 0]^T + \mathbf{T}$, where the additional 0 is the z -axis coordinate of \mathbf{t}_s in SCS. $\mathbf{R} \in \mathbb{R}^{3 \times 3}$ is the rotation matrix and $\mathbf{T} \in \mathbb{R}^{3 \times 1}$ is the translation matrix. They can be derived by geometric calibration, and are usually calibrated once and applied to all gaze targets. The gaze direction \mathbf{g} can be defined by the 3D gaze origin $\mathbf{o} = (x_0, y_0, x_0)$ and the 3D gaze target \mathbf{t}_c . The 3D gaze origin \mathbf{o} is usually defined as the face center or the eye center in CCS, and can be measured by 6-D pose estimation methods. Given \mathbf{o} and \mathbf{t}_c , \mathbf{g} can be calculated as:

$$\mathbf{g} = \frac{\mathbf{t}_c - \mathbf{o}}{\|\mathbf{t}_c - \mathbf{o}\|}$$

Multiple gaze datasets calibrate (\mathbf{R}, \mathbf{T}) through different ways, such as checkerboard, AprilTag (Kellnhofer et al. 2019), calibration ball (Funes Mora, Monay, and Odobez 2014), so there are different $\{(\mathbf{R}'_i, \mathbf{T}'_i)\}_{i=1}^M$, where \mathbf{R}'_i and \mathbf{T}'_i are the actual calibration results with calibration error of the i -th dataset, and M is the number of datasets. The calibration error is fixed in a gaze label acquisition system and different among gaze label acquisition systems, which causes the gaze label deviation in different gaze datasets. Then we analyze the impact of the error in $(\mathbf{R}'_i, \mathbf{T}'_i)$ on the gaze direction \mathbf{g} .

The Impact of Gaze Label Deviation

We construct a typical gaze collection environment for analysis (i.e. the distance between gaze origin \mathbf{o} and gaze target \mathbf{t}_c is 60 cm), the experimental details can be found in the supplementary materials. In this environment, the \mathbf{o} values of different samples remain basically unchanged. Therefore, we make it a constant C to simplify the problem, so the 3D gaze direction $\mathbf{g}' = (p'_x, p'_y, p'_z)$ with deviation can be formulated as:

$$\mathbf{g}' = \frac{\mathbf{R}'\mathbf{t}_c + \mathbf{T}' - C}{\|\mathbf{R}'\mathbf{t}_c + \mathbf{T}' - C\|}$$

Additionally, the rotation matrix \mathbf{R}' can be represented by Euler angles ($pitch, yaw, roll$), the translation matrix $\mathbf{T}' =$

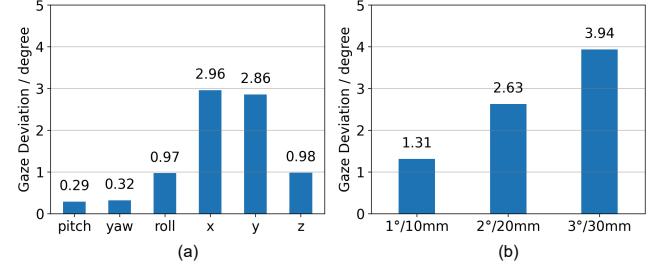


Figure 3: Simulation analysis of the impact of gaze label deviation. (a) is the gaze deviation caused by each single variable of calibration error. (b) is the gaze deviation of coupling multiple calibration error variables.

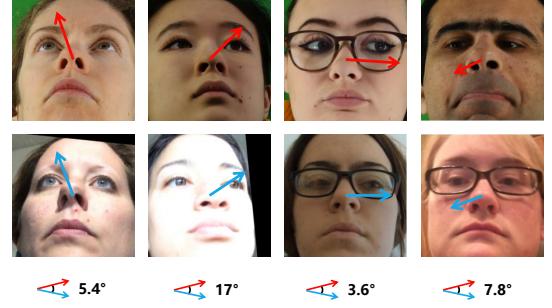


Figure 4: The examples of gaze deviation between gaze datasets. The red and blue arrows are truth gaze labels. In each column, two images from different datasets have similar head poses and eye rotation angles but significantly different gaze directions.

$[x \ y \ z]^T$ denotes the offset on the three coordinate axes. As a result, there are six variables ($pitch, yaw, roll, x, y, z$) totally. We perform sensitivity analysis to analyze the impact of different deviation in $(pitch, yaw, roll, x, y, z)$ on gaze angle separately, then calculate the corresponding gaze deviation. We assume that the calibration errors of multiple gaze acquisition systems follow a Gaussian distribution $X_1 \sim N(\mu, \sigma^2)$ on each variable, where $\mu = 0$ and $\sigma = 1^\circ$ or 1cm. We treat 1° or 1cm as the unit values for analysis, where 1° is for $(pitch, yaw, roll)$, and 1cm is for (x, y, z) . So that the 99.7 % of error is less than 3° or 3cm, and we take 3° or 3cm as typical calibration error values. Then, we add a 3° or 3cm noise on each variable, and calculate the gaze deviation \mathcal{E}_g between \mathbf{g} and \mathbf{g}' by:

$$\mathcal{E}_g = \arccos \left(\frac{\mathbf{g}' \cdot \mathbf{g}}{\|\mathbf{g}'\| \|\mathbf{g}\|} \right)$$

The results are shown in Fig. 3 (a). We notice that the gaze deviation is non-negligible, for example, the deviations caused by x and y are approaching 3° . In fact, the calibration error is the coupling result of multiple variables, so we add a Gaussian distribution noise $X_2 \sim N(v, \tau^2)$ on each variable to simulate the real calibration situation, where $v = 0$, $\tau = 1^\circ$ or 1cm. The Monte Carlo analysis results of gaze deviation within a gaze dataset caused by six variables are

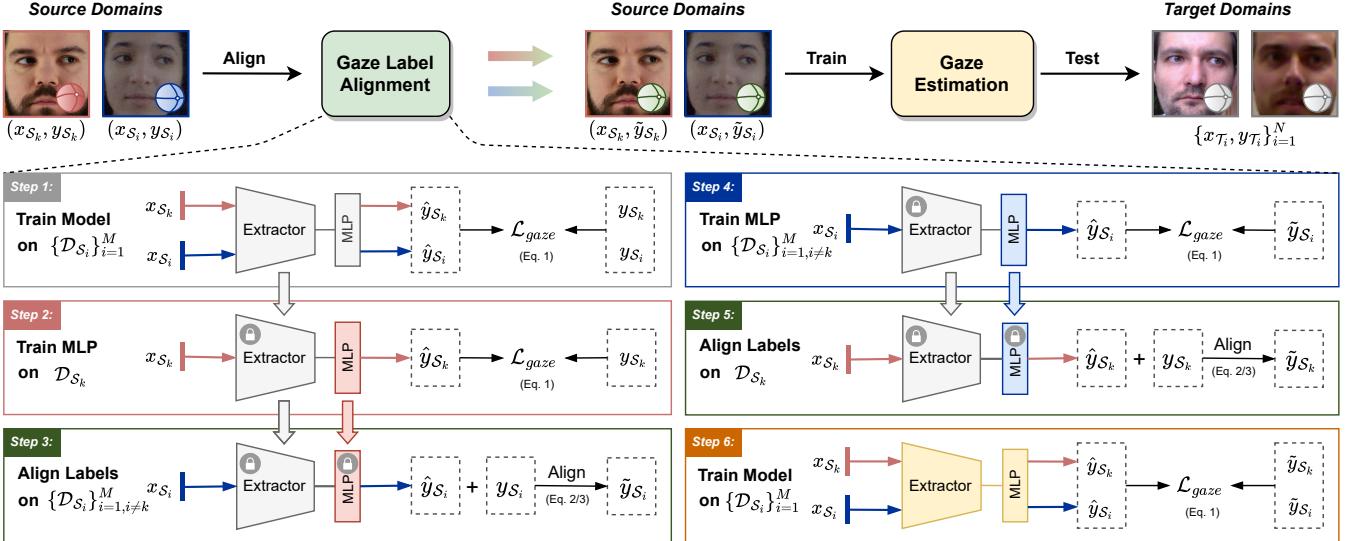


Figure 5: Overview of GLA method. The GLA is the procedure before training gaze estimation model and consists of six steps.

displayed in Fig. 3 (b), at the unit calibration error values, $\mathcal{E}_g = 1.31^\circ$. Further, we change the value of τ . It is found that as the calibration error increases, \mathcal{E}_g also becomes linearly larger.

Actually, we have observed many examples of gaze deviation between multiple gaze datasets, and show them in Fig. 4. The first row of images come from the same gaze dataset, while the second row of images comes from another dataset. Two images in the same column have very similar head poses and eye rotation angles, but with a significant deviation in the truth gaze labels. In a world, the above analysis and phenomenon all illustrate that there are obvious gaze deviations in different gaze datasets.

Gaze Label Alignment

Fig. 5 illustrates the overview of our proposed framework. Given multiple source domains, we first use our proposed gaze label alignment (GLA) method to align the gaze labels. Then the multi-domain data with aligned gaze labels are used to train the gaze estimation model. It is worth noting that, our method can be regarded as the pre-process before training gaze estimation models, and it can be combined with any method to further improve its performance. Suppose we have multiple source domains $\{\mathcal{D}_{S_i}\}_{i=1}^M$, the goal of GLA is to align label distribution among these domains.

As shown in Fig. 5, the GLA method consists of six steps. We first train on all source domains to eliminate data distribution shift. After this, we estimate label distribution gap among domains. If we fix the feature extractor F , and train a gaze regressor on a certain domain \mathcal{D}_{S_k} . Then the predictions of the gaze regressor follow the label distribution of \mathcal{D}_{S_k} . We use this gaze regressor to predict the gaze directions on other domains $\{\mathcal{D}_{S_i}\}_{i=1, i \neq k}^M$, and get the predicted gaze directions \hat{y}_{S_i} corresponding to domain \mathcal{D}_{S_i} . The difference between \hat{y}_{S_i} and ground truth gaze labels y_{S_i} can denote the label deviation between \mathcal{D}_{S_k} and \mathcal{D}_{S_i} . Thus, for

each domain in $\{\mathcal{D}_{S_i}\}_{i=1, i \neq k}^M$ we can fit a function to align y_{S_i} to \hat{y}_{S_i} , and the aligned label \tilde{y}_{S_i} can be considered as the new label of \mathcal{D}_{S_i} without deviation from \mathcal{D}_{S_k} . Similarly, in other domains $\{\mathcal{D}_{S_i}\}_{i=1, i \neq k}^M$, we train a new regressor with the aligned label $\{\tilde{y}_{S_i}\}_{i=1, i \neq k}^M$ to predict \hat{y}_{S_k} , then calculate the aligned labels \tilde{y}_{S_k} . Finally, we retrain the entire gaze estimation model on all source domains with aligned labels to furtherly eliminate the label distribution shift among domains. The detailed description of each step is as follows.

Step1: Eliminate the data distribution shift. Since the distribution shift consists of data distribution $P(X)$ and label distribution $P(Y|X)$, we first need to eliminate data distribution shift. We train gaze estimation model $G = W(F(I))$ on all source domains $\{\mathcal{D}_{S_i}\}_{i=1}^M$, using the convolutional neural networks (CNN) as the feature extractor F and a multi-layer perceptron (MLP) as the gaze regressor W . To avoid the impact of data imbalance, we resample data to maintain the same amount of data for each domain. The following loss function is used to train the model:

$$\mathcal{L}_{gaze}(\hat{y}, y) = \arccos\left(\frac{\hat{y} \cdot y}{\|\hat{y}\| \|y\|}\right) \quad (1)$$

where \hat{y} is the model predicted gaze direction and y is the ground truth label.

Step2: Train gaze regressor on selected domain. We divide all source domains into two parts, the first part is \mathcal{D}_{S_k} and the remaining part is $\{\mathcal{D}_{S_i}\}_{i=1, i \neq k}^M$. Then we fix the the feature extractor F trained on step1 and train a new gaze regressor on \mathcal{D}_{S_k} using the loss function defined in Eq. 1. In this way, we get a gaze regressor W_{S_k} .

Step3: Align labels of remaining domains with selected domain. First, we use the fixed feature extractor F and the gaze regressor W_{S_k} as the gaze estimation model, and evaluate it on remaining domains. For each domain \mathcal{D}_{S_i} in $\{\mathcal{D}_{S_i}\}_{i=1, i \neq k}^M$, we can get the predicted gaze directions \hat{y}_{S_i} . Then, we modify the original truth labels y_{S_i} to align them

with the predicted labels $\hat{y}_{\mathcal{S}_i}$, and generate aligned labels $\tilde{y}_{\mathcal{S}_i}$. Based on the analysis of gaze label deviation, we believe that the calibration error among datasets mainly comes from the gaze target. Hence, we use the following alignment function to calculate $\tilde{y}_{\mathcal{S}_i}$:

$$y_{\mathcal{S}_i} = \frac{\mathbf{t}_{c_i} - \mathbf{o}_i}{\|\mathbf{t}_{c_i} - \mathbf{o}_i\|}, \quad \tilde{y}_{\mathcal{S}_i} = \frac{\tilde{\mathbf{R}}_i \mathbf{t}_{c_i} + \tilde{\mathbf{T}}_i - \mathbf{o}_i}{\|\tilde{\mathbf{R}}_i \mathbf{t}_{c_i} + \tilde{\mathbf{T}}_i - \mathbf{o}_i\|} \quad (2)$$

where \mathbf{t}_{c_i} is the gaze target in CCS of domain $\mathcal{D}_{\mathcal{S}_i}$, and \mathbf{o}_i is the corresponding gaze origin. $\tilde{\mathbf{R}}_i \in \mathbb{R}^{3 \times 3}$ and $\tilde{\mathbf{T}}_i \in \mathbb{R}^{1 \times 3}$ are the fitted parameters of the alignment function, which are shared by all samples and can be estimated by least squares error (LMSE) optimization:

$$\min_{\tilde{\mathbf{R}}_i, \tilde{\mathbf{T}}_i} \mathbb{E}_{(\tilde{y}_{\mathcal{S}_i}, \hat{y}_{\mathcal{S}_i})} [\mathcal{L}_{gaze}(\tilde{y}_{\mathcal{S}_i}, \hat{y}_{\mathcal{S}_i})] \quad (3)$$

where \mathbb{E} is the expectation. Additionally, we can take each person’s data as an alignment unit to calculate the corresponding alignment function parameters to eliminate the impact of individual physiological differences.

Step4: Train gaze regressor on remaining domains. We fix the feature extractor F trained on step1 and train a gaze regressor on $\{\mathcal{D}_{\mathcal{S}_i}\}_{i=1, i \neq k}^N$ with aligned label $\{\tilde{y}_{\mathcal{S}_i}\}_{i=1, i \neq k}^N$. By this way, we obtain a new gaze regressor $W_{\{\mathcal{D}_{\mathcal{S}_i}\}_{i=1, i \neq k}^M}$.

Step5: Align labels of selected domain with remaining domains. In the same way, we use the feature extractor F trained in step1 and the gaze regressor $W_{\{\mathcal{D}_{\mathcal{S}_i}\}_{i=1, i \neq k}^M}$ as gaze estimation model, and evaluate on domain $\mathcal{D}_{\mathcal{S}_k}$. Then modifying the labels $y_{\mathcal{S}_k}$ according to Eq. 2 and Eq. 3 to calculate the aligned labels $\tilde{y}_{\mathcal{S}_k}$.

Step6: Train model with aligned labels. Finally, we use the aligned labels $\{\tilde{y}_{\mathcal{S}_i}\}_{i=1}^M$ to train gaze feature extractor and regressor from scratch on all source domains, then gain a gaze estimation model that alleviates label domain shifts.

Experiments

Data Preparation

Due to the wide range of gaze angle, ETH-XGaze (Zhang et al. 2020) and Gaze360 (Kellnhofer et al. 2019) are commonly used as training datasets. Additionally, we add Gaze-Capture (Krafcik et al. 2016) to construct more source domains. To maintain consistency with previous works (Cheng and Bao 2022; Wang et al. 2022b; Cai et al. 2023), we select MPIIFaceGaze (Zhang et al. 2017) and EyeDiap (Funes Mora, Monay, and Odobez 2014) as our target datasets. Respectively, we denote them as \mathcal{D}_E (ETH-XGaze), \mathcal{D}_G (Gaze360), \mathcal{D}_C (GazeCapture), \mathcal{D}_M (MPIIFaceGaze) and \mathcal{D}_D (EyeDiap). For a detailed description, see the supplementary material.

Implementation Details

We employ ResNet-18 (He et al. 2016) as our gaze feature extractor F , a two layers MLP regressor W to predict gaze angle. For baseline, we employ SGD Optimizer with a learning rate of 5×10^{-2} . The batch size is 126, and all images are resized to 224×224 . We train the gaze estimation

Source	\mathcal{D}_M		\mathcal{D}_D		Avg.	
	w/o	w	w/o	w	w/o	w
\mathcal{D}_E	7.98	6.83	7.65	7.38	7.82	7.11
\mathcal{D}_C	6.72	6.61	7.91	7.44	7.32	7.03
\mathcal{D}_G	8.02	7.62	8.33	7.55	8.18	7.59
$\mathcal{D}_E + \mathcal{D}_C$	6.83	6.58	7.29	6.93	7.06	6.76
$\mathcal{D}_E + \mathcal{D}_G$	6.77	6.63	7.39	7.27	7.08	6.95
$\mathcal{D}_G + \mathcal{D}_C$	5.42	5.05	6.34	5.95	5.88	5.50
$\mathcal{D}_E + \mathcal{D}_G + \mathcal{D}_C$	5.28	4.89	5.91	5.71	5.60	5.30

Table 1: Cross-domain evaluation performances of models trained with different source domains. Results are reported by angular error ($^\circ$), bold denotes the best result on one specific task among each source domain combination. Avg. represents the average error of two cross-domain tasks. w/o indicates without GLA method, and w indicates with GLA method. The bold dataset means that it is selected as \mathcal{D}_{S_k} .

model for 30 epochs utilizing a Cosineannealing LR scheduler (Loshchilov and Hutter 2016) with a 3 epoch warm-up. We perform a data augmentation family with a random color field and greyscale like (Wang et al. 2022b).

Compared Methods

For baseline, we only use Eq. 1 to train the model. For domain generation (DG) methods, we choose PureGaze (Cheng and Bao 2022), CDG (Wang et al. 2022b) and CLIP-Gaze (Yin et al. 2024c) for comparison. The domain adaptation (DA) methods include PnP-GA (Liu et al. 2021b), RUDA (Bao et al. 2022), CSA (Wang et al. 2022b) and UnReGA (Cai et al. 2023). For each method with GLA, we first use baseline method to generate aligned labels $\{\tilde{y}_{\mathcal{S}_i}\}_{i=1}^M$, and retrain gaze model with $\{\tilde{y}_{\mathcal{S}_i}\}_{i=1}^M$. More implementation details can be found in the supplementary material.

Effectiveness of GLA on Multi-Source Domains

In order to verify the effectiveness of GLA on multi-source domains training. We train the baseline models without or with GLA method on different source domains respectively, and conduct cross-domain evaluation. The training source domains are different combinations of \mathcal{D}_E , \mathcal{D}_G and \mathcal{D}_C . When applying GLA method in the single source domain training, we randomly divide the training set into two subsets, and then perform GLA operation. The test datasets are \mathcal{D}_M and \mathcal{D}_D , and the evaluation results are reported in Table 1. Note that the bold dataset means that it is selected as \mathcal{D}_{S_k} .

We can see that, the method with GLA outperforms the method without GLA regardless of on single training domain, two training domains or three training domains. On single training domain, GLA can eliminate the individual deviation, and on multiple domains, GLA can not only eliminate individual deviation but also the deviation of data acquisition systems among different domains.

It is worth noting that the model without GLA trained on $\mathcal{D}_E + \mathcal{D}_C$ has a higher error on \mathcal{D}_M than the model trained on \mathcal{D}_C (6.83° vs 6.72°). We conjecture that the label gap between \mathcal{D}_E and \mathcal{D}_C has led to a higher test error. When we

Task	Methods	\mathcal{D}_M		\mathcal{D}_D		Avg.	
		w/o	w	w/o	w	w/o	w
DG	Baseline	6.83	6.58	7.39	6.93	7.11	6.75
	PureGaze	6.42	6.28	6.77	6.34	6.59	6.31
	CDG	6.63	6.34	6.67	6.42	6.65	6.38
	CLIP-Gaze	6.36	6.13	6.45	6.30	6.40	6.22
UDA	PnP-GA	6.24	5.77	6.32	6.22	6.28	6.00
	RUDA	6.03	5.60	6.07	5.75	6.05	5.68
	CSA	5.79	5.58	6.14	6.00	5.97	5.79
	UnReGA	5.27	5.05	5.86	5.71	5.56	5.38
SDA	Finetuning	5.17	4.62	5.69	5.40	5.43	5.01

Table 2: Performances of GLA method embedded into SOTA methods.

apply GLA in training to eliminate the label shift, the error of the model trained with $\mathcal{D}_E + \mathcal{D}_C$ becomes lower than the model trained with \mathcal{D}_C (6.58° vs 6.61°). With GLA, it is guaranteed that the multi-source domains trained models are always better than the single domain trained models, and more training domains can further improve the performance.

Plugging into the SOTA Methods

Our proposed GLA method can reduce the label gap between different source domains and generate the aligned gaze labels, which is complementary with current methods. By training on the aligned labels instead of the original gaze labels, we can plug the GLA method into SOTA methods to further improve the performance of gaze estimation model. Specifically, we use $\mathcal{D}_E + \mathcal{D}_C$ as source domains, and generate aligned gaze labels through the baseline method with GLA. Then, we train various SOTA methods on the aligned labels, and evaluate the models on \mathcal{D}_M and \mathcal{D}_D . The experimental results are shown in Table 2.

For domain generalization (DG) tasks, we compare the performance of PureGaze, CDG, and CLIP-Gaze on the original labels (w/o GLA) and aligned labels (w GLA). It can be seen that each DG method (w/o GLA) performs better than the baseline method as they reduce the deviation between data distributions. GLA can further decrease the errors on the basis of DG methods, as it can further eliminate the deviation between label distributions. For domain adaptation tasks, we train unsupervised domain adaptation (UDA) methods PnP-GA, RUDA, CSA and UnReGA, and supervised domain adaptation (SDA) method with the original labels (w/o GLA) or aligned labels (w GLA) on the source domains. We randomly choose 100 samples from target domain and report average results of 20 repeated trials. It can be seen that plugging GLA can further reduce estimation errors, achieving a relative improvement of 3-8%, specifically 5% (on PnP-GA), 6% (on RUDA), 3% (on CSA), 3% (on UnReGA), and 8% (on fine-tuning).

The above experiments demonstrate that GLA can be utilized with DG and DA methods to further improve the performance of gaze estimation, which indicates the GLA method is a general and effective method.

Methods	\mathcal{D}_M	\mathcal{D}_D	Avg.
Baseline	6.83	7.39	7.11
GLA _{offset}	6.78	7.11	6.95
GLA _{linear}	6.69	7.04	6.87
GLA _{affine}	6.65	<u>6.97</u>	6.81
GLA _{homography}	<u>6.60</u>	6.99	<u>6.80</u>
GLA _{RT}	6.58	6.93	6.76

Table 3: Performances comparison of different multi-source domain alignment methods.

Source	\mathcal{D}_M			\mathcal{D}_D			Avg.		
	base.	pers.	ds.	base.	pers.	ds.	base.	pers.	ds.
$\mathcal{D}_E + \mathcal{D}_C$	6.83	6.58	<u>6.75</u>	7.29	6.93	<u>7.17</u>	7.06	6.76	<u>6.96</u>
$\mathcal{D}_E + \mathcal{D}_G$	6.77	6.63	<u>6.70</u>	7.39	7.27	<u>7.30</u>	7.08	6.95	<u>7.00</u>
$\mathcal{D}_G + \mathcal{D}_C$	5.42	5.05	<u>5.31</u>	6.34	5.95	<u>6.13</u>	5.88	5.50	<u>5.72</u>
$\mathcal{D}_E + \mathcal{D}_G + \mathcal{D}_C$	5.28	4.89	<u>5.14</u>	5.91	5.71	<u>5.82</u>	5.60	5.30	<u>5.48</u>

Table 4: Performance comparison of different label alignment units. Bold and underline denotes the best and the second best result in a specific task among each source domain combination.

Ablation Study

Ablation Study on Alignment Functions When performing label alignment, we can use different alignment functions to calculate the aligned labels, so we conduct ablation study on alignment functions, these GLA method variants are described as follows.

- GLA_{offset}: calculating the offset between \hat{y} and y , and mapping the original labels y to aligned labels \hat{y} . The alignment function is: $\hat{y} = y + \mathbf{B}$, where $y = (\theta, \phi)$ is the polar equivalent representation of gaze direction \mathbf{g} . $\mathbf{B} \in \mathbb{R}^{1 \times 2}$ is a constant.
- GLA_{linear}: it corresponds the alignment function is: $\hat{y} = \mathbf{K} \cdot y + \mathbf{B}$, where $\mathbf{K} \in \mathbb{R}^{1 \times 2}$ and $\mathbf{B} \in \mathbb{R}^{1 \times 2}$ are linear parameters.
- GLA_{affine}: it corresponds the alignment function is: $\hat{y} = \mathbf{A} \cdot y + \mathbf{T}$, where $\mathbf{A} \in \mathbb{R}^{2 \times 2}$ and $\mathbf{T} \in \mathbb{R}^{1 \times 2}$ are affine transformation parameters.
- GLA_{homography}: it corresponds the alignment function is: $[\tilde{\theta}, \tilde{\phi}, 1]^T = \mathbf{H} \cdot [\theta, \phi, 1]^T$. where $\mathbf{H} \in \mathbb{R}^{3 \times 3}$ is homography transformation parameters.
- GLA_{RT}: it represents the method using Eq. 2.

We use $\mathcal{D}_E + \mathcal{D}_C$ as source domains. The experimental results are shown in Table 3. It can be seen that various alignment functions can achieve good performance improvement compared with the baseline without GLA. among which GLA_{RT} is better than other alignment methods. This also proves that our analysis of gaze label deviation and definition of alignment function are correct and effective.

Ablation Study on Alignment Units We conduct comparative experiments to explore appropriate alignment basic units. Specifically, we treat the data from the same person as a alignment unit or the whole dataset as a unit to align.

Methods	\mathcal{D}_{S_k}	$\{\mathcal{D}_{S_i}\}_{i=1, i \neq k}^M$	\mathcal{D}_M	\mathcal{D}_D	Avg.
Baseline	-	-	6.83	7.39	7.11
$\mathcal{D}_E + \mathcal{D}_G + \mathcal{D}_C$	\mathcal{D}_E	$\mathcal{D}_G + \mathcal{D}_C$	<u>4.89</u>	5.71	5.30
	\mathcal{D}_G	$\mathcal{D}_E + \mathcal{D}_C$	4.84	5.82	<u>5.33</u>
	\mathcal{D}_C	$\mathcal{D}_E + \mathcal{D}_G$	5.17	<u>5.79</u>	5.48

Table 5: Performances comparison of different source domain as \mathcal{D}_{S_k} .

$\mathcal{D}_E + \mathcal{D}_C$ are training domains. The experimental results are shown in Table 4. *base*. in the table represents the baseline method without GLA, *pers.* denotes treating each person’s data as a single domain to align, and *ds.* denotes taking each dataset as an alignment unit. The results show that for each combination of source domains, the method with alignment on dataset level achieves better performance than the one without alignment, and alignment on the person level performs the best. It demonstrates that there exist both systematic deviation between different data acquisition systems and individual deviation between different persons simultaneously. Therefore, it is better to treat the personal data as a separate alignment unit when applying GLA.

Ablation Study on \mathcal{D}_{S_k} Selection When there are multiple source domains, we can select different domain as \mathcal{D}_{S_k} . Therefore, we use $\mathcal{D}_E + \mathcal{D}_G + \mathcal{D}_C$ as source domains, and treat \mathcal{D}_E , \mathcal{D}_G and \mathcal{D}_C as \mathcal{D}_{S_k} to align labels respectively.

The experimental results are shown in Table 5. The test errors of GLA training models with different domain as \mathcal{D}_{S_k} are more than 1.63° lower than the baseline training models. The error difference between different \mathcal{D}_{S_k} is within 0.18° , which is much smaller than the performance benefits brought by GLA (0.18° vs 1.63°). Consequently, although different \mathcal{D}_{S_k} may result in slight performance changes, GLA can still bring stable performance improvement.

Ablation Study on Alignment Times We retrain the model with aligned labels to reduce the domain gap of label distribution between multiple source domains. Therefore, we explore the impact of different label alignment times on model performance. The process described in Fig. 5 is referred to as one complete alignment. When GLA is performed multiple times, each complete alignment results in aligned labels, which are used as the original labels for the next alignment operation. We use $\mathcal{D}_E + \mathcal{D}_C$ as source domains, and test on \mathcal{D}_M and \mathcal{D}_D . The experimental results are shown in Table 6. It can be seen that increasing the number of alignment times does not bring a significant improvement in performance (the average errors are 6.76° , 6.72° and 6.74° respectively). Therefore, one alignment is enough.

Visualization Analysis

We visualize the scatter plot of the estimated $\hat{\theta}, \hat{\phi}$ (X-axis) and ground truth θ, ϕ (Y-axis) of different models in Fig. 6. The first row corresponds to the results of models w/o GLA, and the second row shows the results of models w GLA respectively. All models are trained on $\mathcal{D}_E + \mathcal{D}_G + \mathcal{D}_C$ and predict on a specific person of target domain \mathcal{D}_M . We can find

Times	\mathcal{D}_M	\mathcal{D}_D	Avg.
Baseline	6.83	7.39	7.11
GLA $\times 1$	6.58	6.93	6.76
GLA $\times 2$	6.60	6.89	6.72
GLA $\times 3$	<u>6.59</u>	6.90	6.74

Table 6: Performances comparison of different label alignment times.

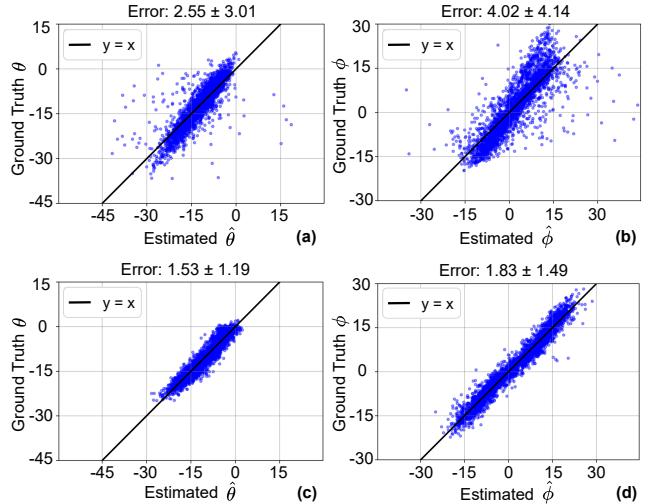


Figure 6: The scatter plot of estimated $\hat{\phi}$ (X-axis) and ground truth ϕ (Y-axis) of different models on a specific person of \mathcal{D}_M . (a) and (b) are the results of the model without GLA. (c) and (d) are the results of the model with GLA.

that the scatters of models w/o GLA are more divergent, and the scatters of models w GLA are more clustered. At the same time, models w/o GLA have higher errors than models w GLA (θ : 2.55° vs 1.53° , ϕ : 4.02° vs 1.83°). This shows GLA can effectively alleviate the label distribution shift and reduce the model prediction uncertainty.

Conclusion

In this paper, we have identified the label distribution shift as a significant factor that affects the generalization error in gaze estimation. To address this issue, we have proposed the Gaze Label Alignment (GLA) algorithm, which aims to align the label distributions among different domains. It is easy to combine GLA with SOTA gaze estimation methods. Through extensive experiments, we have demonstrated the effectiveness of our GLA method. When combined with SOTA domain generalization and adaptation methods, it can also lead to significant improvements in performance.

Acknowledgements

This work was sponsored by National Key R&D Program of China (2023YFE0204200).

References

- Bao, Y.; Liu, Y.; Wang, H.; and Lu, F. 2022. Generalizing Gaze Estimation With Rotation Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4207–4216.
- Biswas, P.; et al. 2021. Appearance-based gaze estimation using attention and difference mechanism. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3143–3152.
- Cai, X.; Zeng, J.; Shan, S.; and Chen, X. 2023. Source-Free Adaptive Gaze Estimation by Uncertainty Reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22035–22045.
- Chang, Z.; Matias Di Martino, J.; Qiu, Q.; Espinosa, S.; and Sapiro, G. 2019. Salgaze: Personalizing gaze estimation using visual saliency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.
- Cheng, Y.; and Bao, Y. 2022. Puregaze: Purifying gaze feature for generalizable gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 436–443.
- Cheng, Y.; Huang, S.; Wang, F.; Qian, C.; and Lu, F. 2020. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 10623–10630.
- Funes Mora, K. A.; Monay, F.; and Odobez, J.-M. 2014. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*, 255–258.
- Guestrin, E. D.; and Eizenman, M. 2006. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering*, 53(6): 1124–1133.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; and Torralba, A. 2019. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6912–6921.
- Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; and Torralba, A. 2016. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2176–2184.
- Li, H.; Wang, Y.; Wan, R.; Wang, S.; Li, T.-Q.; and Kot, A. 2020. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in neural information processing systems*, 33: 3118–3129.
- Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; and Tao, D. 2018. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, 624–639.
- Liu, G.; Yu, Y.; Mora, K. A. F.; and Odobez, J.-M. 2019. A differential approach for gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 43(3): 1092–1099.
- Liu, H.; Nie, H.; Zhang, Z.; and Li, Y.-F. 2021a. Anisotropic angle distribution learning for head pose estimation and attention understanding in human-computer interaction. *Neurocomputing*, 433: 310–322.
- Liu, Y.; Liu, R.; Wang, H.; and Lu, F. 2021b. Generalizing Gaze Estimation with Outlier-guided Collaborative Adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.
- Murphy-Chutorian, E.; and Trivedi, M. M. 2010. Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness. *IEEE Transactions on intelligent transportation systems*, 11(2): 300–311.
- Shah, S. M.; Sun, Z.; Zaman, K.; Hussain, A.; Shoaib, M.; and Pei, L. 2022. A driver gaze estimation method based on deep learning. *Sensors*, 22(10): 3959.
- Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; and Yu, P. 2022a. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.
- Wang, Y.; Jiang, Y.; Li, J.; Ni, B.; Dai, W.; Li, C.; Xiong, H.; and Li, T. 2022b. Contrastive Regression for Domain Adaptation on Gaze Estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19354–19363.
- Xu, M.; Wang, H.; and Lu, F. 2023. Learning a Generalized Gaze Estimator from Gaze-Consistent Feature. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3): 3027–3035.
- Yin, P.; Wang, J.; Dai, J.; and Wu, X. 2024a. Nerf-gaze: A head-eye redirection parametric model for gaze estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2760–2764. IEEE.
- Yin, P.; Wang, J.; Zeng, G.; Xie, D.; and Zhu, J. 2024b. LG-Gaze: Learning Geometry-aware Continuous Prompts for Language-Guided Gaze Estimation. In *Proceedings of the European conference on computer vision (ECCV)*, 1–17. Springer.
- Yin, P.; Zeng, G.; Wang, J.; and Xie, D. 2024c. CLIP-Gaze: Towards General Gaze Estimation via Visual-Linguistic Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6729–6737.
- Zhang, C.; Yao, R.; and Cai, J. 2018. Efficient eye typing with 9-direction gaze estimation. *Multimedia Tools and Applications*, 77: 19679–19696.
- Zhang, X.; Park, S.; Beeler, T.; Bradley, D.; Tang, S.; and Hilliges, O. 2020. ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 365–381. Springer.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2299–2308.