Hybrid-Domain Adaptative Representation Learning for Gaze Estimation

Qida Tan[1], **Hongyu Yang**[2], **Wenchao Du**[2†]

[1] National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, China
[2] College of Computer Science, Sichuan University, Chengdu, China
tanqida@stu.scu.edu.cn, {yanghongyu, wenchaodu.cs}@scu.edu.cn

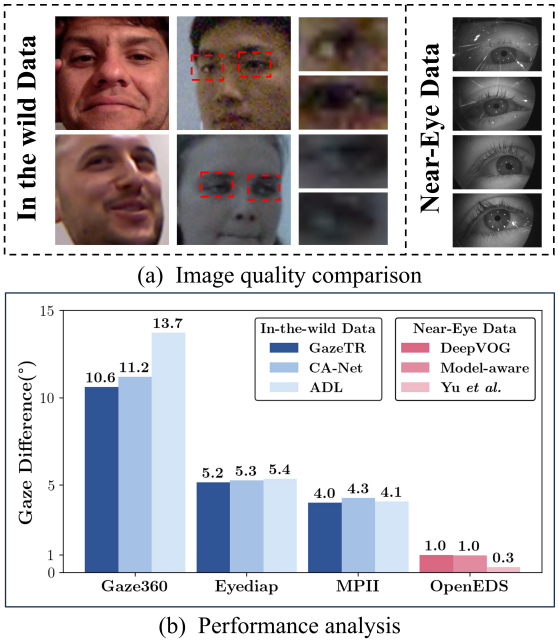
Abstract

Appearance-based gaze estimation, aiming to predict accurate 3D gaze direction from a single facial image, has made promising progress in recent years. However, most methods suffer significant performance degradation in cross-domain evaluation due to interference from gaze-irrelevant factors, such as expressions, wearables, and image quality. To alleviate this problem, we present a novel Hybrid-domain Adaptative Representation Learning (shorted by HARL) framework that exploits multi-source hybrid datasets to learn robust gaze representation. More specifically, we propose to disentangle gaze-relevant representation from low-quality facial images by aligning features extracted from high-quality near-eye images in an unsupervised domain-adaptation manner, which hardly requires any computational or inference costs. Additionally, we analyze the effect of head-pose and design a simple yet efficient sparse graph fusion module to explore the geometric constraint between gaze direction and head-pose, leading to a dense and robust gaze representation. Extensive experiments on EyeDiap, MPIIFaceGaze, and Gaze360 datasets demonstrate that our approach achieves state-of-the-art accuracy of **5.02**°, **3.36**°, and **9.26**° respectively, and present competitive performances through cross-dataset evaluation.

Code — https://github.com/da60266/HARL

Introduction

Gaze estimation, aiming to determine where someone is looking toward or visual attention is located, is a crucial clue for understanding human behaviors, and offers significant assistance for many practical applications, e.g., human-system interaction(Steil, Huang, and Bulling 2018), mental fatigue detection(Lengenfelder et al. 2023) and AR/VR systems(Bao et al. 2023). Existing methods can be approximately divided into two categories: geometry-based and appearance-based. The former focuses on the traditional image processing technologies and geometric gaze model to calculate the gaze direction, which only adapt to near-eye conditions and requires expensive hardware for high-resolution eyeball capturing. With the development of deep learning technologies, appearance-based methods achieve remarkable progress on gaze direction regression from the single facial image in recent years. However, a critical challenging lie in that the eye area only occupies a very small part of the whole face, which leads to the insufficient gaze cues due to lower resolution, e.g., geometric shapes and profiles of pupil and iris, as shown in Fig 1(a). Furthermore, facial expressions, illumination and so on, also degrades the image quality, and leads to limited prediction accuracy and poor generalization on cross-domain evaluation.



(a) Image quality comparison

(b) Performance analysis

Figure 1: (a) Sample quality comparison between in-the-wild and near-eye datasets. (b) Performance comparisons between in-the-wild (Cheng and Lu 2022; Cheng et al. 2020; Kellnhofer et al. 2019) and near-eye datasets(Yiu et al. 2019; Popovic et al. 2023; Feng et al. 2022) respectively.

Recently, all kinds of methods have been explored to alleviate the above problems, including feature distance minimization (Wang et al. 2022), disentangling representation (Sun et al. 2021; Yin et al. 2024), and data generation

†Corresponding author

(Ververas et al. 2024). However, a key point is often overlooked in them, which is that degraded eyeball appearances limit high-quality gaze representation, leading to poor generalization in cross-object gaze estimation. As illuminated in Fig 1(b), high-resolution near-eye images result in the significantly higher prediction accuracy, even with a simple deep network. Therefore, exploiting the high-quality near-eye data to guide the model extracting gaze-relevant representation from low-quality facial image is valuable. However, directly applying existing domain-adaptation methods to aligning them in latent space is unrealistic due to the great domain gap, where the high- and low-quality monocular images are unpaired. Moreover, monocular gaze labels are always lost for low-quality facial images.

Inspired by the unsupervised domain-adaptation (UDA) regression, in this paper, we attempt to introduce high-quality monocular images to disentangle gaze-relevant representation from low-quality facial image with unsupervised learning. We present a novel Hybrid-domain Adaptative Representation Learning (HARL) framework that exploits the labeled high-quality near-eye data to extract monocular gaze representation from unlabeled low-quality facial images in a UDA manner. Specifically, HARL aligns the inverse Gram matrix of the hybrid-domain features to capture inner correlations, which is simple yet efficient and hardly requires any extra computation costs during training and inference. Furthermore, we also consider the effect of head-pose, and construct a sparse-graph fusion module to explore the latent geometry constraints between monocular gaze and head-pose representations, which leads to a dense and robust gaze representation. In short, the main contributions of the paper are summarized:

1) Propose an end-to-end gaze representation learning framework, i.e. HARL, which integrates the idea of UDA into the general appearance learning architecture to extract dense and robust gaze representation from low-quality facial images. *As far as we know, it is the first UDA framework to disentangle gaze representation from hybrid-domain data,* and presents superior performances on in-domain and cross-domain evaluation.

2) Design a simple yet effective sparse-graph fusion module that explores the inherent geometric constraints between the monocular and pose features, and leads to the dense and robust facial gaze representation.

3) Extensive experiments demonstrate that the proposed HARL achieves state-of-the-art gaze accuracy of **3.36**$^\circ$, **5.02**$^\circ$ and **9.26**$^\circ$ on MPIIFaceGaze, EyeDiap and Gaze360 benchmarks, respectively, and also presents competing performances on cross-domain evaluations without any computational costs.

Related Work

Appearance-based Gaze Estimation

The early approaches focus on reconstructing the geometric structure of eyeball, which generally rely on the image processing technologies to locate the boundaries of pupil and iris. Therefore, the deep-learning-based eye segmentation is explored to support more accurate gaze prediction (Yiu et al. 2019; Popovic et al. 2023). However, these methods achieved remarkable accuracy but also require personal calibration and dedicated devices such as depth sensors, infrared cameras and lights. Appearance-based approaches directly estimate gaze vector from the facial image captured by the web camera, which builds an end-to-end mapping between the image and the gaze label (Kellnhofer et al. 2019; Wang et al. 2023; O Oh, Chang, and Choi 2022). Therefore, these methods have made great progress in recent years. However, a crucial challenge lies that the gaze representation is sensitive to facial appearances, which limits the prediction accuracy and generalization ability of the model. Leading works focus on exploring domain-adaptation (DA) (Bao et al. 2022; Liu et al. 2021; Cai et al. 2023) and domain-generalization (DG) (Xu, Wang, and Lu 2023; Bao and Lu 2024a,b; Xu and Lu 2024) gaze estimation. The former aims to align gaze representations between the source and target domains, which generally requires to accessing the source and target domain data. Instead, the latter directly learn robust gaze representation from source domain data only. Although all of them achieved some improvements on cross-datasets evaluation, complex networks and carefully-designed training strategies leads to expensive computation costs, which is still unsuitable in practical applications.

Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) aims to adapt the model to target domain with unlabeled samples, which is widely applied to classical computer vision tasks (Rahman et al. 2020; Vu et al. 2019). The general methods focus on feature alignment by adversarial learning or explicit losses such as maximum mean discrepancy (MMD) (Long et al. 2017; Rahman et al. 2020), which have been applied to gaze estimation (Guo et al. 2020; Wang et al. 2022). However, these methods ignored a crucial problem that the robust gaze representation is difficult to acquire in the source domain due to the undesired degradation. Instead, our approach aims to disentangle the monocular gaze representation via UDA, which exploits the high-quality data as the source domain, low-quality facial data as the target domain, to capture feature correlations. It brings significant gains for in-dataset and cross-dataset evaluation.

Graph Neural Networks

Graph neural networks (GNNs) have received tremendous attention in causal reasoning from structured and non-structured data (Wu et al. 2020). The major component of the GNNs is the node feature aggregation technique, with which node can update its weight by interacting with other nodes (Kipf and Welling 2017). Meanwhile, they adopt the same strategy in aggregating the information from different feature dimensions. However, inspired by recent advances on GNNs, there are potential benefits to treat the dimensions differently during the aggregation process (Jin et al. 2021). Considering that extracting dense and robust gaze representations from low-quality images is challenging even with powerful supervision, thus, we investigate to enable heterogeneous contributions of feature dimensions with GNNs, which aims to explore fine-grained feature correlations for

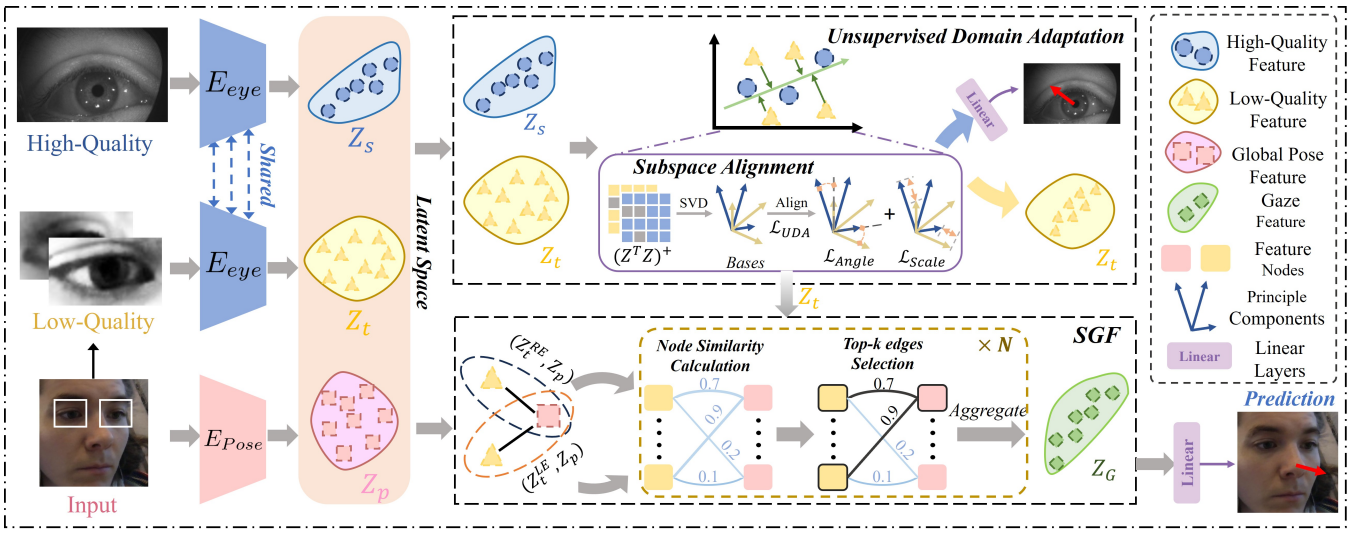


Figure 2: Overview of our framework. The proposed method explores a dual branch network to learning monocular and pose representation, a sparse-graph fusion module is used to generate dense gaze representation.

robust gaze representation. *As far as we know, this is also the first work to exploit GNNs in gaze estimation.*

Proposed Method

Given a training sample (I, g) drawn from specific domain $\mathcal{D}$, where I denotes the input facial image, g is the corresponding gaze label, which is generally expressed by the Euler angle (pitch and yaw) (ϕ, ψ). The appearance-based gaze estimation is defined as:

$$Z = E(I; \Theta), \hat{\mathcal{G}} = R(Z; \beta), \tag{1}$$

where $E(\cdot)$ is an encoder, which aims to extract gaze-relevant representation Z from input, $R(\cdot)$ represents a linear regressor that estimate the gaze vector $\hat{\mathcal{G}}$. Generally, E and R are coupled into an end-to-end appearance learning architecture. Θ and β denote the learnable parameters.

In order to learn robust gaze representation Z, most appearance-based approaches, including supervised and unsupervised-based, focus on exploring more powerful encoder E to minimize the distribution difference between source feature Z_s and target feature Z_t, where the samples from source and target domains are all degraded facial images. In contrast, our method takes the high-resolution near-eye data as the source domain, the low-quality facial images are used as target domain. The key to it is aligning Z_t to Z_s to ensure effective monocular gaze representation learning, and then the head-pose factor is also considered with a novel sparse-graph fusion module, which leads to a dense and robust facial gaze representation. The proposed framework is shown in Fig 2, which exploits a dual-branch encoding architecture, and contains two main modules, i.e. a Unsupervised Domain Adaptation Learning (UDAL) module, a Head-Pose injected Sparse-Graph Fusion (SGF) module.

Unsupervised Domain Adaptation Learning Monocular gaze direction is a critical clue to infer facial gaze estimation. However, estimating it from low-quality facial images is challenging due to lack of monocular gaze labels and clear eye-appearance context. Therefore, UDAL exploits the idea of UDA to disentangle monocular gaze representation from low-quality facial images, which takes the high-quality monocular near-eye data as source domain, and leverages the pseudo-inverse low-rank property to align the scale and angle in a selected subspace generated by the pseudo-inverse Gram matrix of the two domains, leading to a disentangling representation learning framework.

Assuming that a linear regressor R with parameter β is utilized to estimate the monocular gaze directions $\hat{\mathcal{G}}$ in Eq 1, i.e. $\hat{\mathcal{G}} = Z\beta$, which has an ordinary least squares (OSL) closed-form solution,

$$\hat{\beta} = \left(Z^T Z\right)^{-1} Z^T \hat{\mathcal{G}} \tag{2}$$

where $\left(Z^T Z\right)^{-1} \in \mathbb{R}^{n \times n}$ is the inverse of the Gram Matrix. $Z^T \hat{\mathcal{G}} \in \mathbb{R}^{n \times 2}$ projects gaze feature to label space.

Suppose that we have a set of low- and high-quality near-eye image features with the same gaze direction. Our goal is to ensure that the low-quality image features would produce the same gaze direction through a shared R, thereby achieving consistent constraints between low- and high-quality image features, i.e. Z_t and Z_s. This objective is formulated as:

$$\begin{aligned} &\text{If } \hat{\mathcal{G}}_s = \hat{\mathcal{G}}_t \\ &\text{then } \left(Z_s^T Z_s\right)^{-1} Z_s^T \hat{\mathcal{G}}_s = \left(Z_t^T Z_t\right)^{-1} Z_t^T \hat{\mathcal{G}}_t, \end{aligned} \tag{3}$$

where $\hat{\mathcal{G}}_s$ and $\hat{\mathcal{G}}_t$ denote the gaze vectors from source and target images. When both them produce the same gaze prediction through a shared $R(\cdot; \beta)$, it implies that the encoder $E(\cdot)$ can extract consistent gaze-relevant features from different domains. By achieving it, *low-quality gaze features Z_t can be aligned with high-quality features Z_s, ensuring that the gaze representations remain highly relevant and contain*

rich gaze information. Returning to the Eq 3, it could be implemented by aligning $(Z^TZ)^{-1}Z^T$. Inspired by the recent advances on the UDA in regression (Chen et al. 2021; Nejjar, Wang, and Fink 2023), we can achieve the alignment of $(Z^TZ)^{-1}Z^T$ by incorporating angular and scale constraints on the inverse Gram matrix $(Z^TZ)^{-1}$ during training, which not only ensures the alignment of the inverse Gram matrix but also contributes to a well-aligned Z.

The Gram Matrix Z^TZ of the feature can be decomposed using singular value decomposition (SVD) (Gloub and Van Loan 1996):

$$\left(Z^TZ\right) = \left(UDV^T\right)^T \left(UDV^T\right) = V\Lambda V^T, \tag{4}$$

where the orthogonal matrix $V \in \mathbb{R}^{n\times n}$ is identical to the matrix in the SVD of Z and $\Lambda \in \mathbb{R}^{n\times n}$ is the diagonal matrix containing the squared eigenvalues of Z. Given a feature matrix $Z \in \mathbb{R}^{b\times n}$ where n is always greater than b, the corresponding gram matrix has rank $r \leq b$. As a result, the Gram matrix is not full-rank and therefore not invertible. In such cases, the Moore-Penrose-pseudo-inverse (Ben-Israel and Greville 2006) can be used to generalize the concept of the matrix inverse and provide a stable solution for further computations involving matrix inversion. Given the ordered eigenvalues of the $\left(Z^TZ\right) \in \mathbb{R}^{n\times n}$, $\lambda_1 \geq ... \geq \lambda_r \geq ... \geq \lambda_n$, the pseudo-inverse of $\left(Z^TZ\right)$ can be expressed as:

$$\begin{aligned} G^+ &= \left(Z^TZ\right)^+ = V\Lambda^+V^T \\ &= V \left(\begin{array}{ccc|c} \frac{1}{\lambda_1} & & & 0 \\ & \ddots & & \vdots \\ & & \frac{1}{\lambda_r} & 0 \\ \hline 0 & \cdots & 0 & 0 \end{array}\right) V^T. \end{aligned} \tag{5}$$

The r-principal components derived from the the Λ can be regarded as the subspace bases. Then we can measure the principle angles of source and target spaces using the following equation:

$$\cos\left(\theta_i^{s\leftrightarrow t}\right) = \frac{G_{i,s}^+ \cdot G_{i,t}^+}{||G_{i,s}^+|| \cdot ||G_{i,t}^+||}. \tag{6}$$

The cosine similarity between the spans of the source and target subspaces is given by $M = [\cos\left(\theta_1^{s\leftrightarrow t}\right), ..., \cos\left(\theta_n^{s\leftrightarrow t}\right)]$. We aim to make the angles between the two subspaces as close as possible, leading to the following angle alignment constraint:

$$\mathcal{L}_{Angle}\left(Z_s, Z_t\right) = ||\mathbb{I} - M||_1, \tag{7}$$

where $\mathbb{I}$ is the vector of 1 with shape of n. The scale alignment is regularized by minimizing the difference between the r-principal eigenvalues,

$$\mathcal{L}_{Scale}\left(Z_s, Z_t\right) = ||\ \lambda_{s,i=1,..,r} - \lambda_{t,i=1,..,r}||_2. \tag{8}$$

Finally, the Unsupervised Domain-Adaptation (UDA) loss is the combination of these two items,

$$\mathcal{L}_{UDA} = \mathcal{L}_{Angle}\left(Z_s, Z_t\right) + \mathcal{L}_{Scale}\left(Z_s, Z_t\right). \tag{9}$$

As shown in Fig 2, in practice, we first crop the eye regions from the facial image and combine them with high-quality monocular near-eye images as input to the shared encoder E_{eye}, it hardly requires any extra computational costs.

Sparse-Graph Fusion Previous works (Yue et al. 2024; Liang, Bao, and Lu 2025) have explored the effect of head-pose for gaze estimation, it heavily affects the prediction accuracy in the wild. Existing methods directly model a global appearance-based network to extract pose prior from the facial image, which always leads to poor generalization due to the irrelevant interference (e.g., facial texture and expression). To this end, we also consider the effect of the head-pose but exploit a more simple and effective manner. We take a pretrained dense facial landmark model to extract robust pose representations, which provides rich pose and gaze clues by landmarks locations. This process is defined as

$$Z_p = E_{Pose}(I;\theta), \tag{10}$$

where the θ is fixed model parameter and I denotes the input image. In practice, we use the output of the penultimate layer of the model as dense pose representation Z_p.

Considering the robust facial gaze representation is always low-ranked and dense. After extracting binocular gaze features $Z_t = \{Z_t^{LE}, Z_t^{RE}\}$ and pose feature Z_p, where Z_t^{LE} and Z_t^{RE} denote the feature embeddings from the left- and right-eye images separately, we further explore the SGF module to capture the inner geometric constraints among them, while aggregating highly-relevant gaze information.

Supposing binocular features $\{Z_t^{LE}, Z_t^{RE}\}$ and facial pose feature Z_p are viewed as the single node, separately, it is too simple to capture effective gaze features with three nodes and two edges in the resulting graph. Moreover, although the representations are low-dimensional and dense, they remain affected by interference-induced redundancy. Inspired by the efficient feature gating mechanism (Jin et al. 2021) that dynamically adjusts the contribution of each feature dimension during aggregation to enhance the impact of the important features, we further construct a subgraph between the monocular and pose features, where each feature dimension in them is treated as an individual node, and edges are defined based on feature similarity among each dimension. A two-layer MLP network is employed to compute the node similarity between monocular and facial pose features in a self-adaptive manner:

$$\mathcal{S}_{i,j} = Linear(\sigma(Linear(n_i - n_j^p))) \tag{11}$$

where σ denotes a ReLU activation. In addition, n and n^p denote the nodes from binocular gaze and facial pose representations, respectively.

We select the top-k pose feature nodes that related to each monocular feature node in Z_t. The adjacent matrix $\mathcal{AD}$ of all nodes in the single subgraph is constructed as

$$\mathcal{AD}_{i,j} = \begin{cases} 1, & \text{if } j \in \text{top-}k\{\mathcal{S}_{i,k} \mid k = 1, \ldots, q\}, \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

where q indicates the number of dimensions.

To capture inherent geometric relationship for each node, we iteratively update node features by aggregating neighboring nodes at each layer,

$$\begin{aligned} n_i^{l+1} &= Linear(n_i^l) + \sum_{j\in\mathcal{N}_i} Linear(n_j^{(l,p)}) \\ \text{where} \quad \mathcal{N}_i &= \{j \mid \mathcal{AD}_{i,j} = 1\}, \end{aligned} \tag{13}$$

where $\mathcal{N}_i$ denotes a set of neighborhood nodes with center node n_i, and the l indicates the current layer of GNNs.

After aggregating node features through several GNNs layers, SGF output a unified full-face gaze representation Z_G, the final gaze vector regression is implemented by stacking two linear layers.

Joint Optimization The proposed HARL is an end-to-end learning framework, the total loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{MSE}(\mathcal{G}_{Face}, \hat{\mathcal{G}}_{Face}) + \mathcal{L}_{MSE}(\mathcal{G}_{eye}, \hat{\mathcal{G}}_{eye}) + \lambda \mathcal{L}_{UDA}(Z_s, Z_t) \quad (14)$$

where the mean square error (MSE) is used to supervise gaze vector regression. $\hat{\mathcal{G}}_{Face}$ and $\hat{\mathcal{G}}_{eye}$ denote predicted gaze vector for facial and near-eye images separately. Z_s and Z_t are the extracted monocular gaze representations from high- and low-quality monocular images respectively. λ is a hyperparameter to balance the effect of UDA loss.

Experiments

Experimental Setup

Datasets. We select three benchmarks to evaluate the performance of HARL, i.e. Eyediap (Funes Mora, Monay, and Odobez 2014), MPIIFace (Zhang et al. 2019) and Gaze360 (Kellnhofer et al. 2019). Eyediap has $16k$ images captured in a controlled laboratory environment with screen and floating targets. We divide it into four clusters and apply four-fold cross-validation for in-dataset tests. MPIIFace Consists of $45k$ images captured by webcams during daily laptop usage. We perform leave-one-subject-out cross-validation for in-dataset tests. Gaze360 comprises $101k$ images collected using a 360° camera in outdoor street settings. Considering our HARL depends on monocular gaze representations from left and right-eye images, we filtered out samples with missing eye regions due to large pose during the training. Furthermore, we use the OpenEDS2020 (Palmero et al. 2021) as the high-quality source data, which contains $180k$ high-resolution eye images captured using head-mounted VR/AR devices. We only select a subset for training.

Implementation Details. We use the ResNet18 (He et al. 2016) as the basic encoder to implement the UMGRL, and an extra linear layer is exploited to regress monocular gaze directions for high-quality near-eye data. For SGF module, we select the PiPNet (Jin, Liao, and Shao 2020) as the pre-trained pose encoder, which also used the ResNet18 as the basic encoder. Then we implement our sparse-graph network with 4 GNN layers. We use the SGD optimizer with learning rate $lr = 1e^{-4}$ during training. High-resolution monocular images from OpenEDS2020 are center-cropped and resized to 224×224. Low-quality monocular images are also cropped from the facial images, and then are resized and grayscaled. The hyperparameter λ in Eq 14 is set 0.5. We train HARL with 50, 30, and 30 epochs on EyeDiap, MPIIFace, and Gaze360 respectively for in-domain evaluation. For cross-domain evaluation, we use Gaze360 as the source domain, and train the model with 20 epochs only.

For in-domain evaluation, we select some representative methods for comparison, including ADL (Kellnhofer et al. 2019), SAtten-net (O Oh, Chang, and Choi 2022), GazeCaps (Wang et al. 2023), 3DGazeNet (Ververas et al. 2024), IEH (Yue et al. 2024), L2CS-NET (Abdelrahman et al. 2023), GazeTR (including GazeTR-Pure and GazeTR-Hybrid) (Cheng and Lu 2022), CA-Net (Cheng et al. 2020) and GFNet (Hu and Huang 2023). For cross-domain evaluation, we also select representative Domain-Adaptation methods and Domain-Generalization (DG) methods for comprehensive comparison, including PnP-GA (Liu et al. 2021), DAGEN (Guo et al. 2020), ADDA (Tzeng et al. 2017), CRGA (Wang et al. 2022), GVBGD (Cui et al. 2020), RUDA (Bao et al. 2022), Full-Face (Zhang et al. 2017), RT-Gene (Fischer, Chang, and Demiris 2018), CA-Net (Cheng et al. 2020), GazeTR (Cheng and Lu 2022), GazeCaps (Wang et al. 2023), ADL (Kellnhofer et al. 2019), PureGaze (Cheng, Bao, and Lu 2022), CLIP-Gaze (Yin et al. 2024), GazeCon (Xu, Wang, and Lu 2023), AGG (Bao and Lu 2024a), and GLA(Zeng et al. 2025). Moreover, The Mean Angular Error (MAE) is used as the common evaluation metric.

Method	MPIIFace	EyeDiap	Gaze360
ADL	4.06°	5.36°	13.73°
SAtten-net	4.04°	5.25°	10.70°
GazeCaps	4.06°	<u>5.10°</u>	10.40°
3DGazeNet	4.0°	-	<u>9.60°</u>
IEH	<u>3.49°</u>	-	9.89°
L2CS-NET	3.92°	-	10.40°
GazeTR-Pure	4.74°	5.72°	13.58°
GazeTR-Hybrid	4.00°	5.17°	10.62°
CA-Net	4.27°	5.27°	11.20°
GFNet	3.96°	5.40°	-
HARL	**3.36°**	**5.02°**	**9.26°**

Table 1: Mean Angular Error (MAE) results of different methods. The best and second-best results are **bolded** and <u>underlined</u>, respectively.

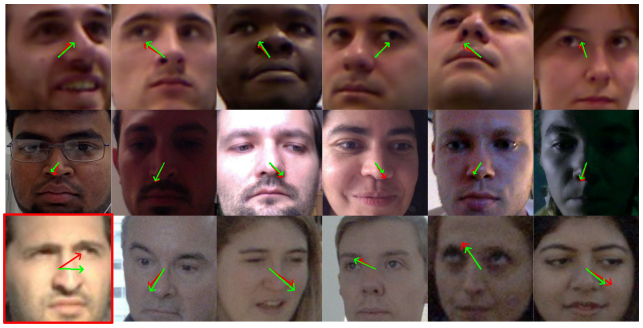


Figure 3: Visualized results from different datasets. The red and green arrows denote the predicted and ground truth gaze vector, separately.

Comparison with SOTA Methods

In-Domain Evaluation We first perform in-domain evaluation on three benchmarks, and the quantitative results are listed in Tab 1. Most methods have achieved some improve-

Methods	Type	Source	Target	G → E	G → M	Avg.
PnP-GA	DA	Yes	100	7.92°	6.18°	7.05°
DAGEN	DA	Yes	500	12.90°	8.74°	10.8°
ADDA	DA	Yes	500	12.90°	6.61°	9.76
CRGA	DA	Yes	100	6.68°	**6.09°**	6.39°
GVBGD	DA	Yes	1000	12.44°	7.64°	10.0°
RUDA	DA	Yes	100	**5.86°**	6.20°	**6.03°**
Full-Face	DG	No	0	14.42°	11.13°	12.8°
RT-Gene	DG	No	0	38.60°	21.81°	30.2°
CA-Net	DG	No	0	31.41°	27.13°	29.3°
GazeTR	DG	No	0	8.88°	7.96°	8.42°
GazeCaps	DG	No	0	9.20°	9.20°	9.20°
ADL	DG	No	0	11.86°	11.36°	11.6°
PureGaze	DG	No	0	9.32°	9.28°	9.30°
CLIP-Gaze$^-$	DG	No	0	7.73°	7.55°	7.74°
CLIP-Gaze	DG	No	0	**7.06°**	**6.89°**	**6.98°**
GazeCon	DG	No	0	8.52°	7.82°	8.17°
AGG	DG	No	0	7.93°	7.87°	7.90°
GLA	DG	No	0	7.55°	7.62°	7.59°
HARL	DG	No	0	7.63°	7.49°	7.56°

Table 2: MAE results for cross-domain evaluations. 'Type' denotes whether the method belongs to domain-adaptation (DA) or domain-generalization (DG). 'Source' indicates whether the source domain sample is required during testing. 'Target' denotes whether the target-domain sample is required and the number of samples from the target domain. The 'G→E' and 'G→M' represents using Gaze360 as source domain, EyeDiap and MPIIFace are viewed as target domains, respectively.

ments on MPIIFace and EyeDiap, e.g., IEH (Yue et al. 2024) also explored the transformer blocks to fuse local monocular and global facial gaze representation, which decreased the MAE to $3.49°$ on MPIIFace. GazeCaps (Wang et al. 2023) exploited the capsule network with a self-attention-routing mechanism, and achieved the competing result with $5.10°$ on EyeDiap. Instead, our HARL achieves significantly better MAE results with $3.36°$ and $5.02°$ on MPIIFace and EyeDiap. Moreover, there is a clear tendency for all methods to experience significant performance degradation on Gaze360, mainly due to its more unconstrained data collection environment, which leads to greater degradation. Our method also achieves the best MAE with clear gains against 3DGazeNet (Ververas et al. 2024). It supports that our HARL could capture reliable gaze representations for low-quality facial images. Fig 3 gives the visualized results on typical real scenes, our method presents better robustness for lighting, expression, and pose. Note that the red rectangle denotes a special sample with an error gaze label, but HARL still predicts the correct gaze direction.

Cross-Domain Evaluation Tab 2 lists the detailed results on the two cross-domain evaluation tasks. It is clear that the DA-based method achieved significantly better MAE results than the DG-based methods. However, they always require source and target samples to fine-tune the model, which is impossible in practical scenes. DG-based methods present competing performances, although with slightly higher MAEs, e.g., CLIP-Gaze (Yin et al. 2024) achieved the best MAE results with $7.06°$ and $6.89°$ on EyeDiap and MPIIFace benchmarks separately compared to other methods. However, CLIP-Gaze requires explicit text prompts corresponding to the facial image, e.g., expressions, illuminations, poses, and glasses. Removing personalized prompts, i.e. CLIP-Gaze$^-$, the performance drops a lot. In contrast, our HARL hardly requires any extra computational costs during the training and testing, which significantly improves the generalization ability with competing MAEs, making it more suitable for real-world applications.

Variant	EyeDiap	G → M	G → E	#Params
Baseline	$5.54°$	$11.97°$	$12.56°$	–
w/ $\mathcal{L}_{UDA}$	$5.42°$	$9.81°$	$9.49°$	+ 0.M
w/ E_{Pose}	$5.35°$	$9.69°$	$10.42°$	+ 0.M
w/ *SGF*	$5.47°$	$9.48°$	$11.08°$	+ 2.1M
w/ $\mathcal{L}_{UDA}$ + E_{Pose}	$5.20°$	$8.08°$	$8.48°$	+ 0.M
w/ E_{Pose} + *SGF*	$5.23°$	$8.35°$	$8.65°$	+ 2.1M
w/ $\mathcal{L}_{UDA}$ + *SGF*	$5.17°$	$8.27°$	$9.45°$	+ 2.1M
Full Model	**5.02°**	**7.49°**	**7.63°**	+ 2.1M

Table 3: MAE results from each variant. The 'G→M' and 'G→E' denote using Gaze360 as source domain for training, MPIIFace and EyeDiap are viewed as target domains for evaluation, respectively.

Strategy	w/o SubGraph	top-3	top-2	top-1
MAE	$5.55°$	$5.22°$	$5.13°$	**5.02°**

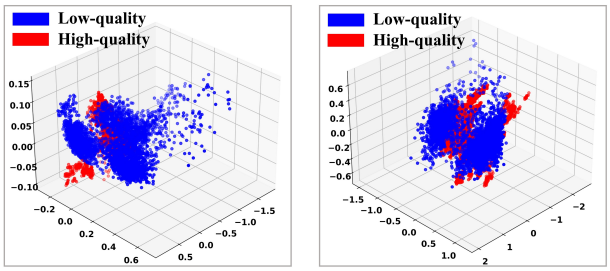
Table 4: MAE results on EyeDiap with different graph construction strategies.

Ablation Study

In this section, we mainly analyze the effect of key components in our HARL, including UDAL (i.e. $\mathcal{L}_{UDA}$), E_{Pose} and SGF module, experiments are performed on EyeDiap and MPIIFace to evaluate their effects under in-domain and cross-domain settings.

Component Analysis Our baseline exploits a dual branch network to encoder monocular gaze and global pose representations separately, which are simply implemented with two ResNet-18 networks, the outputs are concatenated along channel dimension and then fed into a two-layer MLP to predict final gaze direction. Note that the pretrained pose network also used the ResNet-18 as the backbone. Then we insert the $\mathcal{L}_{UDA}$, E_{Pose}, and SGF into the baseline, where the $\mathcal{L}_{UDA}$ and E_{Pose} hardly increase any extra parameters and computational costs due to the shared encoder.

Tab 3 lists the detailed results. The baseline present some ability on in-domain evaluation, but it has poor generalization on cross-domain tasks. Integrating $\mathcal{L}_{UDA}$, E_{Pose} and SGF into baseline all bring some gains for in-domain and cross-



(a) w/o $\mathcal{L}_{UDA}$ (b) w/ $\mathcal{L}_{UDA}$

Figure 4: Monocular gaze feature distribution visualization. Different colors denote features from different samples.

domain evaluation, but w/ $\mathcal{L}_{UDA}$ leads to significantly better generalization performances. Furthermore, the variants with different combinations among them all bring clear performance gains compared to the variants with single module only. Finally, our full model (i.e. HARL) achieves the most significant performance improvements across in-domain and cross-domain evaluations with few parameters.

Sparse Graph Construction For SGF module, we also explore the effect of sparse subgraph construction, the results are shown in Tab 4. Firstly, removing subgraph in each layer (i.e., the monocular gaze features and pose feature are viewed as individual node) leads to a more simple graph with three nodes and two edges only, it results in the obvious performance drop, i.e. inner geometric constraints among them are not exploited well. Furthermore, we explore the sparse edge connection in subgraph, which relies on adjacent matrix $\mathcal{AD}$ in Eq 12. We select different top-k to construct subgraph, and observe an interesting phenomenon that using fewer similar nodes to construct edges brings the better performances, which leads to an extremely sparse subgraph in each layer of SGF. This also significantly enhances the inference efficiency of the model.

Visualization Analysis We visualize monocular gaze representation from UDAL to evaluate the effectiveness of the $\mathcal{L}_{UDA}$, where we only remove it during the training, but low-quality and high-quality monocular images share an encoder. The features visualized results are shown in Fig 4. It is obvious that removing $\mathcal{L}_{UDA}$ leads to significant distribution gap in embedding space. Instead, with $\mathcal{L}_{UDA}$, the features from high- and low-quality samples are aligned with consistent angles and scales, which implies the effectiveness of our domain-adaptation constraints.

To analyze the final gaze features of different variants, we further visualize the distribution of gaze features on domain generalization task $G \rightarrow M$ with t-SNE. Results are shown in Fig 5, where the feature points with close gaze direction shared with similar colors. For the baseline model from Tab 3, as shown in Fig 5(a), the features with different gaze directions are mixed together and the feature cluster is quite dispersed. After injecting the E_{Pose} and SGF module into the baseline (shown in Fig 5(b)), the feature distributions tend to become ordered, but they are still mixed. Then,

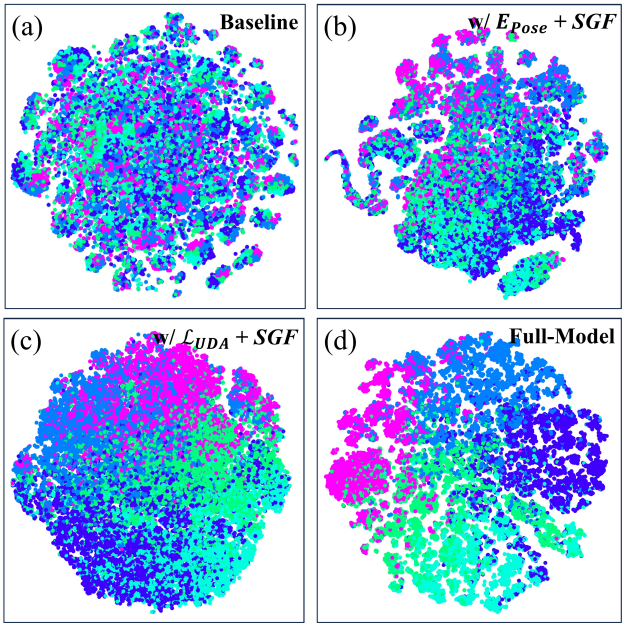


Figure 5: Visualization of the full-face gaze feature distribution. Different colors denotes different gaze directions and close gaze direction share similar colors.

we introduce $\mathcal{L}_{UDA}$ and SGF module, the feature distribution becomes ordered with obvious cluster (shown in Fig 5(c)). The best feature cluster appears in our full model, as shown in Fig 5(d), the gaze direction and feature similarities have a strong correlation, which supports our insight, disentangling gaze representations from unconstrained facial images.

Conclusion

In this paper, we present a simple yet efficient Hybrid-domain Adaptative Representation Learning (HARL) framework, which is the first method to exploit hybrid-domain data to disentangle monocular gaze representation in an unsupervised domain adaptation manner, and further explores an efficient sparse-graph network to fuse head-pose representation, leading to an adaptative gaze representation learning architecture. The proposed HARL requires few parameters and computational costs and achieves significant gains in both in-domain and cross-domain evaluations. While many concrete implementations of the general idea, including utilizing powerful graph networks and fusion modules, are possible, we show that a simple design already achieves superior results, which provides new insights to solve robust gaze estimation in the wild.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62301345, in part by Sichuan Province Postdoctoral Special Funding under Grant TB2025010, in part by the National Basic Scientific Research Project of China under Grant JCKY2024110C080.

References

Abdelrahman, A. A.; Hempel, T.; Khalifa, A.; Al-Hamadi, A.; and Dinges, L. 2023. L2cs-net: Fine-grained gaze estimation in unconstrained environments. In *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, 98–102. IEEE.

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.

Bao, Y.; and Lu, F. 2024a. From Feature to Gaze: A Generalizable Replacement of Linear Layer for Gaze Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1409–1418.

Bao, Y.; and Lu, F. 2024b. Unsupervised Gaze Representation Learning from Multi-view Face Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1419–1428.

Bao, Y.; Wang, J.; Wang, Z.; and Lu, F. 2023. Exploring 3d interaction with gaze guidance in augmented reality. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 22–32. IEEE.

Ben-Israel, A.; and Greville, T. N. 2006. *Generalized inverses: theory and applications*. Springer Science & Business Media.

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 (CVPR)*, 22035–22045.

Chen, X.; Wang, S.; Wang, J.; and Long, M. 2021. Representation Subspace Distance for Domain Adaptation Regression. In *International Conference on Machine Learning*.

Cheng, Y.; Bao, Y.; and Lu, F. 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.

Cheng, Y.; and Lu, F. 2022. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 3341–3347. IEEE.

Cui, S.; Wang, S.; Zhuo, J.; Su, C.; Huang, Q.; and Tian, Q. 2020. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12455–12464.

Feng, Y.; Goulding-Hotta, N.; Khan, A.; Reyserhove, H.; and Zhu, Y. 2022. Real-time gaze tracking with event-driven eye segmentation. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 399–408. IEEE.

Fischer, T.; Chang, H. J.; and Demiris, Y. 2018. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European conference on computer vision (ECCV)*, 334–352.

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 (ETRA)*, 255–258. ACM.

Gloub, G. H.; and Van Loan, C. F. 1996. Matrix computations. *Johns Hopkins Universtiy Press, 3rd edtion*.

Guo, Z.; Yuan, Z.; Zhang, C.; Chi, W.; Ling, Y.; and Zhang, S. 2020. Domain adaptation gaze estimation by embedding with prediction consistency. In *Proceedings of the Asian Conference on Computer Vision*.

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.

Hu, D.; and Huang, K. 2023. GFNet: Gaze Focus Network using Attention for Gaze Estimation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2399–2404. IEEE.

Jin, H.; Liao, S.; and Shao, L. 2020. Pixel-in-Pixel Net: Towards Efficient Facial Landmark Detection in the Wild. *International Journal of Computer Vision*, 129: 3174 – 3194.

Jin, W.; Liu, X.; Ma, Y.; Derr, T.; Aggarwal, C.; and Tang, J. 2021. Graph Feature Gating Networks. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, 813–822. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384469.

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.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907*.

Lengenfelder, C.; Hild, J.; Voit, M.; and Peinsipp-Byma, E. 2023. Pilot Study on Gaze-Based Mental Fatigue Detection During Interactive Image Exploitation. In Harris, D.; and Li, W.-C., eds., *Engineering Psychology and Cognitive Ergonomics*, 109–119. Cham: Springer Nature Switzerland.

Liang, Z.; Bao, Y.; and Lu, F. 2025. De-confounded Gaze Estimation. In *European Conference on Computer Vision*, 219–235. Springer.

Liu, Y.; Liu, R.; Wang, H.; and Lu, F. 2021. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3835–3844.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2017. Conditional Adversarial Domain Adaptation. In *Neural Information Processing Systems*.

Nejjar, I.; Wang, Q.; and Fink, O. 2023. DARE-GRAM: Unsupervised domain adaptation regression by aligning inverse gram matrices. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11744–11754.

O Oh, J.; Chang, H. J.; and Choi, S.-I. 2022. Self-attention with convolution and deconvolution for efficient eye gaze

estimation from a full face image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4992–5000.

Palmero, C.; Sharma, A.; Behrendt, K.; Krishnakumar, K.; Komogortsev, O. V.; and Talathi, S. S. 2021. Openeds2020 challenge on gaze tracking for vr: Dataset and results. *Sensors*, 21(14): 4769.

Popovic, N.; Christodoulou, D.; Paudel, D. P.; Wang, X.; and Van Gool, L. 2023. Model-aware 3D Eye Gaze from Weak and Few-shot Supervisions. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 746–751. IEEE.

Rahman, M. M.; Fookes, C.; Baktashmotlagh, M.; and Sridharan, S. 2020. On minimum discrepancy estimation for deep domain adaptation. *Domain Adaptation for Visual Understanding*, 81–94.

Steil, J.; Huang, M. X.; and Bulling, A. 2018. Fixation detection for head-mounted eye tracking based on visual similarity of gaze targets. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, 1–9.

Sun, Y.; Zeng, J.; Shan, S.; and Chen, X. 2021. Cross-encoder for unsupervised gaze representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3702–3711.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7167–7176.

Ververas, E.; Gkagkos, P.; Deng, J.; Doukas, M. C.; Guo, J.; and Zafeiriou, S. 2024. 3DGazeNet: Generalizing 3D Gaze Estimation with Weak-Supervision from Synthetic Views. In *European Conference on Computer Vision*, 387–404. Springer.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2517–2526.

Wang, H.; Oh, J. O.; Chang, H. J.; Na, J. H.; Tae, M.; Zhang, Z.; and Choi, S.-I. 2023. Gazecaps: Gaze estimation with self-attention-routed capsules. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2669–2677.

Wang, Y.; Jiang, Y.; Li, J.; Ni, B.; Dai, W.; Li, C.; Xiong, H.; and Li, T. 2022. Contrastive regression for domain adaptation on gaze estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19376–19385.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Philip, S. Y. 2020. Graph neural networks: A review and applications. *Artificial Intelligence*, 288: 103254.

Xu, M.; and Lu, F. 2024. Gaze from origin: learning for generalized gaze estimation by embedding the gaze frontalization process. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6333–6341.

Xu, M.; Wang, H.; and Lu, F. 2023. Learning a generalized gaze estimator from gaze-consistent feature. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 3027–3035.

Yin, P.; Zeng, G.; Wang, J.; and Xie, D. 2024. CLIP-Gaze: Towards General Gaze Estimation via Visual-Linguistic Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6729–6737.

Yiu, Y.-H.; Aboulatta, M.; Raiser, T.; Ophey, L.; Flanagin, V. L.; Zu Eulenburg, P.; and Ahmadi, S.-A. 2019. DeepVOG: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning. *Journal of neuroscience methods*, 324: 108307.

Yue, J.; Lan, P.; Zhou, Y.; and Dong, Z. 2024. Gaze Estimation by Integrating Eye and Head Representation. In *Proceedings of the 2024 12th International Conference on Communications and Broadband Networking*, 136–141.

Zeng, G.; Wang, J.; Xu, Z.; Yin, P.; Ren, W.; Xie, D.; and Zhu, J. 2025. Gaze label alignment: Alleviating domain shift for gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9780–9788.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 51–60.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2019. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1): 162–175.