De-confounded Gaze Estimation

Ziyang Liang[1,2], Yiwei Bao[1], and Feng Lu[1,2*]

[1] State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University, Beijing, China
[2] Peng Cheng Laboratory, Shenzhen, China
{liangziyang, baoyiwei, lufeng}@buaa.edu.cn

Abstract. Deep-learning based gaze estimation methods suffer from sever performance degradation in cross-domain settings. One of the primary reason is that the gaze estimation model is confounded by gaze-irrelevant factor during estimation, such as identity and illumination. In this paper, we propose to tackle this problem by causal intervention, an analytical tool that alleviates the impact of confounding factors by using intervening the distribution of confounding factors. Concretely, we propose the Feature-Separation-based Causal Intervention (FSCI) framework for generalizable gaze estimation. The FSCI framework first separates gaze features from gaze-irrelevant features. To alleviate the impact of gaze-irrelevant factors during training, the FSCI framework further implements causal intervention by averaging gaze-irrelevant features using the proposed Dynamic Confounder Bank strategy. Experiments show that the proposed FSCI framework outperforms SOTA gaze estimation methods in varies cross-domain settings, improving cross-domain accuracies by up to 36.2% over the baseline and 11.5% over SOTA methods, respectively, without touching target domain data.

Keywords: Gaze estimation · Causal intervention · Domain generalization

1 Introduction

Human gaze direction offers a wealth of information, reflecting the underlying cognitive and emotional status behind human behavior in social environments [29]. Serving as a crucial clue in understanding human actions, gaze estimation has widespread applications in various fields such as virtual/augmented reality [18,30,37], human-computer interaction [15,32,33], healthcare [4,17], and assisted driving [1, 22, 31]. In recent years, leveraging the superior performance of Convolutional Neural Network (CNN) in extracting image features, CNN-based gaze estimation methods demonstrate good performance in the within-dataset tests. However, due to the domain gap in data distribution across different domains, performance of CNN-based methods decreases significantly in

* Corresponding Author.
This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62372019. The code is available at https://github.com/pylzy-98/code-for-De-confounded-Gaze-Estimation.

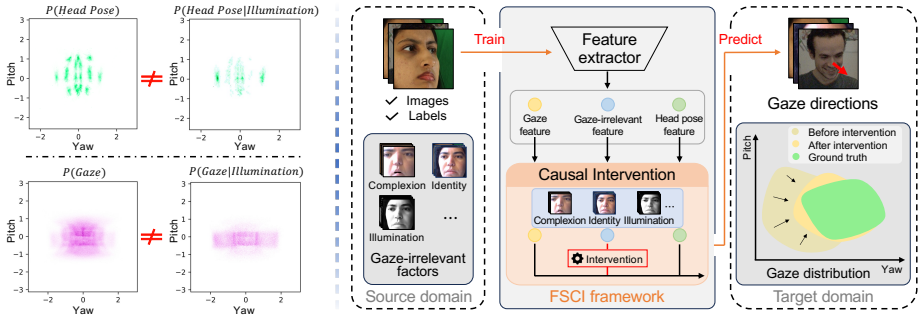

Fig. 1: Left: The head pose and gaze distributions in the ETH-XGaze [41] dataset, showing the difference between the head pose/gaze distribution and their corresponding conditional distributions. **Right:** The overall structure of the proposed Feature-Separation-based Causal Intervention (FSCI) framework. FSCI employs causal intervention to mitigate the influence of gaze-irrelevant feature on gaze estimation.

cross-domain settings. To address the issue, Unsupervised Domain Adaptation (UDA) approaches have been proposed [16,24,34]. These approaches utilize some samples from the target domain and employ unsupervised methods to enable the model to learn the target domain distribution. However, the applicability of UDA is limited, as they require knowledge about the target domain.

More recently, researchers have begun to focus on the Domain Generalization (DG) problem. DG methods aim to enhance the generalization ability of gaze estimation models without utilizing any target domain data. Recent works have proposed to utilize adversarial training [7] and adversarial disturbance [39] for generalizable training. The DG problem is both more practical and challenging, which still requires further exploration.

To study the Domain Generalization problem, we revisit the logic behind tasks in the field of computer vision. *Today's computer vision systems are good at telling us "what" and "where", yet bad at knowing "why"* [35], e.g., why the gaze direction is as it is? Causal intervention can address such questions. A classic example is the influence of temperature (T) on ice cream sales (x) and the number of drowning deaths (y), creating the false impression that x and y are directly proportional. The T confounds the causal relationship between x and y. By intervening on x (for instance, closing all ice cream stores), the information from x does not transfer to y, thereby eliminating the spurious correlation between x and y.

Backing to the above question: "why the gaze direction is as it is?", we try to provide an answer using causal intervention under the DG setting. We dive deep into the cause and effect of the domain gap. Since it is impossible for human-collected datasets to cover all scenarios, there will inevitably be biases in the distribution of gaze-related factors (*e.g.*, head pose and gaze) and gaze-irrelevant factors (*e.g.*, illumination and identity). As a result, the gaze estimation model captures spurious correlations between gaze-irrelevant factors and gaze during source domain training. However, such spurious correlation does not stand in

target domains, since the distribution bias is domain-specific. Thus, the spurious correlation captured by gaze estimation model is the cause of performance degradation in cross-dataset tests.

To address above problem, we propose the Feature-Separation-based Causal Intervention gaze estimation framework (FSCI framework), a Domain Generalization method that improves the generalization ability of gaze estimation models without touching target domain data. The target of the FSCI framework is to learn the true **Causes** of the gaze, *i.e.* estimating gaze from the eye appearance for better generalization. The proposed FSCI framework employs Causal Intervention to prevent the gaze estimation model from being **Confounded** by the spurious correlations caused by distribution bias in source domain. Specifically, we first establish the causal graph of key elements in gaze estimation task, including the input image $\boldsymbol{Z}$, gaze-irrelevant feature $\boldsymbol{I}$, head pose feature $\boldsymbol{H}$, gaze feature $\boldsymbol{G}$ and gaze direction $\boldsymbol{g}$. According to the causal graph, we propose the Feature Separation Module (FSM) to separate gaze feature, head pose feature and gaze-irrelevant features. Then, we propose the Causal Intervention Module (CIM) to alleviate the impact of gaze-irrelevant factors in gaze estimation. By calculating the moving average of the confounder factor $\boldsymbol{I}$, the proposed CIM employs the *do*-calculus $P(\boldsymbol{g}|do(\boldsymbol{I}))$ to alleviate the spurious correlation between the gaze-irrelevant features and gaze during source domain training.

Experiments show that the proposed FSCI framework improves the cross-domain gaze estimation performance significantly, outperfoming State-of-the-Art (SOTA) gaze Domain Generalization methods. Further analysis demonstrates that the proposed FSCI framework alleviates the impact of gaze-irrelevant features from two aspects: (1) the impact of spurious correlation between gaze-irrelevant factors and gaze; (2) the impact of spurious correlation between head pose distribution and gaze distribution generated by the common cause of gaze-irrelevant factors (explained in Sec. 3). The contributions are as follows:

- We propose the Feature-Separation-based Causal Intervention gaze estimation framework (FSCI framework), a gaze domain generalization method that utilizes causal intervention to alleviate the impact of gaze-irrelevant factors during gaze estimation for better generalization ability.
- We introduce Dynamic Confounder Bank, a continuous and dynamic implementation of the *do*-calculus in gaze estimation task. The Dynamic Confounder Bank could potentially benefit other regression tasks.
- Experimental results show that the FSCI framework achieves consistent improvement over four different cross-domain tasks with two different backbone models, ranging from 27.6% to 36.2% , and achieving up to an 11.5% improvement over SOTA methods.. It improves the generalization ability of gaze estimation models significantly without using target domain data.

2 Related work

The appearance-based gaze estimation method overcomes the limitations of early gaze estimation techniques, which required the construction of 3D eye-

ball models. These methods only requires eye images [11, 21, 26, 34, 42], face images [16, 19, 43], or both [2, 8, 20] to estimate gaze direction.

Zhang *et al.* [42] proposed a CNN-based gaze estimation method, leveraging the CNNs' ability to extract image features, marking the first instance of gaze direction estimation from eye images. Zhang *et al.* [43] utilized full-face images as input and applied a CNN with spatial weighting to feature maps, effectively encoding facial images. Cheng *et al.* [9] observed the phenomenon of "two eye asymmetry" in estimating gaze direction using both the left and right eyes. Based on this observation, they proposed an Asymmetric Regression Evaluation Network, which significantly improves the performance of gaze estimation. Although the above methods have good performance in the within-dataset setting, their performance typically degrades significantly in new domains.

2.1 Domain Generalization

Domain generalization tasks require models to be trained on a source domain and then tested for performance in a target domain. Chen *et al.* [6] proposed a Dilated-Net architecture, which employs multiple pooling or downsampling layers to extract better features from eye images. Cheng *et al.* [7] mitigated the impact of gaze-irrelevant factors, such as illumination and identity, on cross-domain gaze estimation by extracting purified gaze features. Xu *et al.* [39] regarded gaze-irrelevant factors as detrimental interference and utilized them to disrupt training data, enabling the model to adapt solely to gaze-related features.

2.2 Unsupervised Domain Adaptation

Unlike DG, UDA tasks require the unsupervised use of samples from the target domain as a crucial means to enhance the model's generalization capability in the target domain. Guo *et al.* [13] proposed a UDA method that can alleviate the impact of inter-personal diversity. Bao *et al.* [3] found that human gaze vectors possess rotational consistency and, based on this property, proposed a UDA method. Wang *et al.* [36] proposed a gaze adaptation method, namely Contrastive Regression Gaze Adaptation, which pulls features corresponding to similar gaze directions closer. However, regardless of the type of UDA method, information from the target domain is required, which undoubtedly increases the limitations on the applicability of UDA methods.

2.3 Causal Inference

In computer vision, causal relationships may exist between different features of images. Unlike simple statistical correlation methods, which cannot address causal relationships, causal inference can effectively tackle this issue. Wang *et al.* [35] proposed an unsupervised feature representation learning method based on causal intervention. Yang *et al.* [40] argued that harmful biases, which can be regarded as confounders, mislead the learning process of models. Chen *et al.* [5] proposed a method to address the issue of spurious correlations between

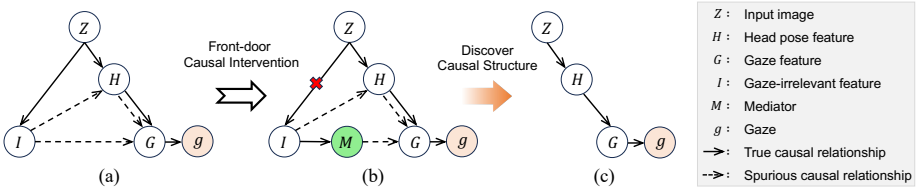

Fig. 2: Illustration of our gaze estimation causal graph. After intervening on $\boldsymbol{Z}$, the path $\boldsymbol{Z} \rightarrow \boldsymbol{I}$ will be blocked.

questions and answers in Visual Question Answering tasks by synthesizing counterfactual samples. In addressing VQA problems, Liu *et al.* [23] considered that some concepts frequently appearing in linguistic and visual modalities should be treated as confounders and proposed a VQA model based on causal inference. In the field of computer vision, causal inference has emerged as a highly effective tool for addressing problems involving causal relationships.

3 Causal Inference in Gaze Estimation

In this section, we first formulate a tailored causal graph for the Gaze Estimation (GE) task in Sec. 3.1. Then, we bridge the gap between theoretical causal graph and specific implementation through a series of formula derivation in Sec. 3.2.

3.1 Causal Graph in Gaze Estimation

Following prior study [28], we formulate the causal graph for the Gaze Estimation task as a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, consisting of variable nodes $\mathcal{N}$ and causal effect links $\mathcal{E}$. In GE tasks, there are five key variables, including the input image $\boldsymbol{Z}$, gaze-irrelevant feature $\boldsymbol{I}$, head pose feature $\boldsymbol{H}$, gaze feature $\boldsymbol{G}$ and gaze direction $\boldsymbol{g}$, as shown in Fig. 2. (a). As for causal effect links, we use solid arrows to denote true causal effects and dashed arrows to denote spurious causal effects. Next, we provide a detailed explanation of the causal graph.

$\boldsymbol{Z} \rightarrow \boldsymbol{H} \rightarrow \boldsymbol{G} \rightarrow \boldsymbol{g}$. The input image content $\boldsymbol{Z}$ determines the head pose of the subject $\boldsymbol{H}$. The head pose of the subject also affects the appearance of the subject's eyes, result in $\boldsymbol{H} \rightarrow \boldsymbol{G}$. Thus, the causal path $\boldsymbol{Z} \rightarrow \boldsymbol{H} \rightarrow \boldsymbol{G} \rightarrow \boldsymbol{g}$ represents the desired true casual effect from the input image to gaze, *i.e.* the gaze direction should be estimated from gaze and head pose feature.

$\boldsymbol{Z} \rightarrow \boldsymbol{I}$. The input image content $\boldsymbol{Z}$ determines the distribution of gaze-irrelevant features $\boldsymbol{I}$ such as illumination, identity, complexion and *etc.* Fundamentally, the distribution of $\boldsymbol{I}$ comes from the collecting procedure and environment of the dataset. The distribution of $\boldsymbol{I}$ is domain-specific since the collecting procedure and environment of each dataset is different.

$\boldsymbol{I} \dashrightarrow \boldsymbol{G}$. Since it is impossible for a dataset to cover all scenarios, the distribution of gaze-irrelevant factors and gaze is inevitably biased. The extracted gaze feature $\boldsymbol{G}$ would be affected by the gaze-irrelevant features $\boldsymbol{I}$ due to the

spurious correlation between them observed by the gaze estimation model during source domain training. For example, in dataset $\mathcal{A}$, if a large number of data is collected during mobile phone usage at night, gaze estimation model trained in dataset $\mathcal{A}$ would tend to produce downward gaze directions on dark images. However, this spurious correlation does not hold in other datasets, leading the model to produce inaccurate estimations in cross-dataset tasks.

$\boldsymbol{I} \dashrightarrow \boldsymbol{H} \dashrightarrow \boldsymbol{G}$. Due to the same reason as $\boldsymbol{I} \dashrightarrow \boldsymbol{G}$, head pose is also affected by gaze-irrelevant factors. Furthermore, because the $\boldsymbol{I}$ is the common cause of $\boldsymbol{H}$ and $\boldsymbol{G}$ (hence $\boldsymbol{I}$ is also known as confounder [27]), the confounder $\boldsymbol{I}$ causes the negative effect of misleading the model to learn spurious correlations between $\boldsymbol{H}$ and $\boldsymbol{G}$. Following the same example in last paragraph, models trained in dataset $\mathcal{A}$ could tends to produce downward gaze directions because subjects usually lower their heads when using mobile phone.

Above analysis demonstrates that the gaze-irrelevant factors affects the gaze estimation process from two paths: $\boldsymbol{I} \dashrightarrow \boldsymbol{G}$ and $\boldsymbol{I} \dashrightarrow \boldsymbol{H} \dashrightarrow \boldsymbol{G}$. Since above spurious correlation is domain-specific, it decreases the accuracy of gaze estimation models in cross-domain tasks. In the next section, we try to alleviate the impact of $\boldsymbol{I}$ by front-door adjustment.

3.2 Causal Intervention via Front-door Adjustment

In causal inference, eliminating the influence of $\boldsymbol{I}$ means computing the distribution $P(\boldsymbol{g}|do(\boldsymbol{I}))$, where $do(\boldsymbol{I})$ denotes the intervention on $\boldsymbol{I}$ to eliminate spurious correlations caused by distribution bias in source domain. For details on the $do(\cdot)$, refer to [27]. Following previous study [23], to apply front-door adjustment, an additional mediator should be inserted between $\boldsymbol{I}$ and $\boldsymbol{G}$, which construct a front-door path $\boldsymbol{Z} \rightarrow \boldsymbol{I} \rightarrow \boldsymbol{M} \dashrightarrow \boldsymbol{G}$, as shown in Fig. 2 (b). By using front-door adjustment [27],

$$\begin{aligned} P(\boldsymbol{g}|do(\boldsymbol{I})) &= \textstyle\sum_{\boldsymbol{m}} P(\boldsymbol{g}|do(\boldsymbol{I}), \boldsymbol{M}=\boldsymbol{m}) P(\boldsymbol{M}=\boldsymbol{m}|do(\boldsymbol{I})) \\ &= \textstyle\sum_{\boldsymbol{m}} P(\boldsymbol{g}|do(\boldsymbol{M}=\boldsymbol{m})) P(\boldsymbol{M}=\boldsymbol{m}|\boldsymbol{I}) \\ &= \textstyle\sum_{\boldsymbol{I}'} \sum_{\boldsymbol{m}} P(\boldsymbol{g}|\boldsymbol{M}=\boldsymbol{m}, \boldsymbol{I}=\boldsymbol{I}') P(\boldsymbol{M}=\boldsymbol{m}|\boldsymbol{I}) P(\boldsymbol{I}=\boldsymbol{I}'). \end{aligned} \tag{1}$$

The above Eq. (1) holds based on the total probability formula and the causal structure after intervention. According to Eq. (1), as it involves computing the series twice, the computational cost of reproducing this process using neural networks would be very huge. However, the Normalized Weighted Geometric Mean (NWGM) [38] can be used to address this issue. Based on Eq. (1), by using NWGM,

$$P(\boldsymbol{g}|do(\boldsymbol{I})) \overset{\text{NWGM}}{\approx} P(\boldsymbol{g}|\mathbb{E}[\boldsymbol{M}|\boldsymbol{I}], \mathbb{E}\boldsymbol{I}). \tag{2}$$

Through front door adjustment, the path $\boldsymbol{Z} \rightarrow \boldsymbol{I}$ is cut off, which mitigate the influence of $\boldsymbol{I}$ on gaze estimation. Subsequently, as shown in Fig. 2.(b), the spurious correlation between $\boldsymbol{H}$ and $\boldsymbol{G}$ is also eliminated, due to $\boldsymbol{I}$ no longer being a common cause of both (the spurious correlation $\boldsymbol{I} \dashrightarrow \boldsymbol{G}$ is eliminated). As shown in Fig. 2(c), the results of the intervention can be reflected in two

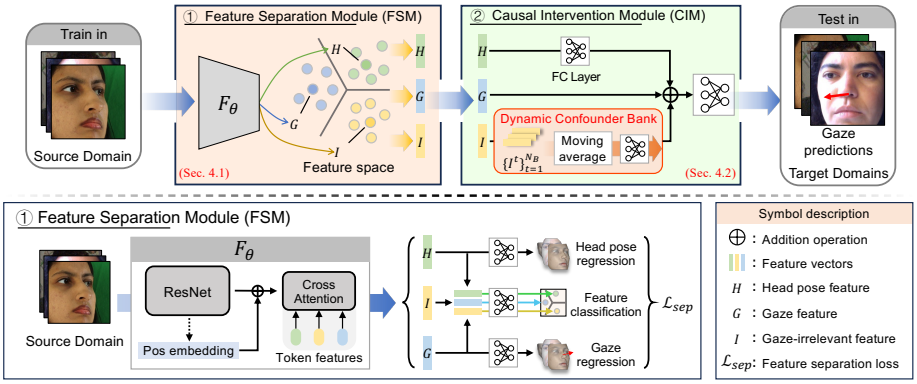

Fig. 3: The overview of the proposed Feature-Separation-based Causal Intervention gaze estimation framework, which consists of two modules: 1) the Feature Separation Module (FSM) and 2) the Causal Intervention Module (CIM). First, the FSM separates the extracted features of the input images into head pose feature, gaze feature, andgaze-irrelevant feature. Then, based on causal intervention, the CIM aggregates the causal effect of gaze-irrelevant feature on gaze direction through the Dynamic Confounder Bank, combined with gaze-related feature to estimate the gaze direction.

aspects, firstly, the spurious correlation between $\boldsymbol{I}$ and $\boldsymbol{G}$ is eliminated, and secondly, the spurious correlation between $\boldsymbol{H}$ and $\boldsymbol{G}$ caused by $\boldsymbol{G} \leftarrow\!\!- \boldsymbol{I} \rightarrow \boldsymbol{H}$ is also eliminated, which is proved through experiments in Sec. 5.4. In Sec. 4, we design a network model corresponding to Eq. (2) to address the negative impact of $\boldsymbol{I}$ in the gaze estimation process.

4 Method

For DG tasks, our framework consists of two modules: the Feature Separation Module (FSM) and the Causal Intervention Module (CIM). To mitigate the influence of $\boldsymbol{I}$ through CIM, we employ the FSM to extract input image features, separating them into three parts: head pose feature, gaze feature, and gaze-irrelevant feature.

4.1 Feature Separation Module

Fig. 3 illustrates the workflow of the FSM. Initially, the cross-attention layer separates the information from different channels, extracted by the ResNet [14] convolutional layer, into three distinct feature vectors: $\boldsymbol{H}$, $\boldsymbol{G}$, and $\boldsymbol{I}$. During the training phase, to ensure that the information among $\boldsymbol{H}$, $\boldsymbol{G}$, and $\boldsymbol{I}$ does not overlap, we classify them using a single-layer fully connected layer. Moreover, to align the information in $\boldsymbol{H}$ and $\boldsymbol{G}$ with head pose feature and gaze feature, respectively, we regress $\boldsymbol{H}$ and $\boldsymbol{G}$ using different single-layer fully connected layers. The FSM loss function is as follows:

$$\begin{aligned}\mathcal{L}_{sep}(\boldsymbol{H},\boldsymbol{G},\boldsymbol{I},\boldsymbol{h}_l,\boldsymbol{g}_l) = \mathcal{L}_1(\phi_1(\boldsymbol{H}),\boldsymbol{h}_l) + \mathcal{L}_1(\phi_2(\boldsymbol{G}),\boldsymbol{g}_l) \\ +\mathcal{L}_{CE}(\phi_3(\{\boldsymbol{H},\boldsymbol{G},\boldsymbol{I}\}),\{l_H,l_G,l_I\}),\end{aligned} \tag{3}$$

where $\mathcal{L}_1(\cdot,\cdot)$ is $\mathcal{L}_1$ loss function, and $\mathcal{L}_{CE}$ is Cross Entropy loss function. $\phi_1(\cdot), \phi_2(\cdot)$, and $\phi_3(\cdot)$ represent different fully connected layers, respectively. l_h, l_g, and l_I are the one-hot classification labels corresponding $\{\boldsymbol{H}, \boldsymbol{G}, \boldsymbol{I}\}$. $\boldsymbol{h}_l$ and $\boldsymbol{g}_l$ represent the ground truth label of head pose and gaze, respectively. During the training phase, we employ an optimizer to separately optimize the parameters of the FSM.

4.2 Causal Intervention Module

Fig. 3 illustrates the process of CIM, which deconfounds the features separated by the FSM through causal intervention. Assume that the feature space is $\mathbb{R}^n$. Eq. (2) only provides the probability of $[\boldsymbol{g}|do(\boldsymbol{I})]$. However, the value of $[\boldsymbol{g}|do(\boldsymbol{I})]$ is required in the GE. From the perspective of probability measure, Eq. (2) can be understood as estimating the gaze direction $\boldsymbol{g}$ under specific conditions and then using the probability measure function P to map the estimated value of $\boldsymbol{g}$ to a probability. Inspired by [35], we parameterize a network to obtain $[\boldsymbol{g}|do(\boldsymbol{I})]$,

$$[\boldsymbol{g}|do(\boldsymbol{I})] \approx \left[\boldsymbol{g}|\mathbb{E}[\boldsymbol{M}|\boldsymbol{I}], \mathbb{E}\boldsymbol{I}\right] = \boldsymbol{W}_1\mathbb{E}\boldsymbol{I}, \tag{4}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{n\times n}$ is a learnable parameter. From causal graph Fig. 2 (b), we can see that $\boldsymbol{M}$ is only influenced by $\boldsymbol{I}$, which means the information of $\mathbb{E}[\boldsymbol{M}|\boldsymbol{I}]$ can be derived from $\mathbb{E}\boldsymbol{I}$. $\boldsymbol{W}_1\mathbb{E}\boldsymbol{I}$ indicates that, by mapping $\mathbb{E}\boldsymbol{I}$, the model can learn the causal effect of $\mathbb{E}[\boldsymbol{M}|\boldsymbol{I}]$ and $\mathbb{E}\boldsymbol{I}$ to $\boldsymbol{g}$. However, $[\boldsymbol{g}|do(\boldsymbol{I})]$ only represents the causal effect of $\boldsymbol{I}$ to $\boldsymbol{g}$. In the input images, head pose information and gaze information are the primary sources of gaze-related information. Therefore, we aggregate these features to estimate gaze direction $\boldsymbol{g}$:

$$\boldsymbol{g} \approx \phi(\boldsymbol{W}_1\mathbb{E}\boldsymbol{I} + \boldsymbol{G} + \boldsymbol{W}_2\boldsymbol{H}), \tag{5}$$

where $\boldsymbol{W}_2 \in \mathbb{R}^{n\times n}$ is a learnable parameter. $\boldsymbol{W}_2\boldsymbol{H}$ represents mapping the head pose feature into the head-pose-related gaze feature, which serves as supplementary information of gaze-related features to assist gaze estimation. $\phi(\cdot)$ denotes a two-layer fully connected network used to regress this gaze-related information to the gaze direction $\boldsymbol{g}$. The loss function of CIM is $\mathcal{L}_1$ loss. Next, we present the process of the Dynamic Confounder Bank.

Dynamic Confounder Bank. To alleviate the impact of factors causing spurious correlations, a common strategy in previous work is to construct a confounder dictionary [23,35,40]. However, this strategy is not applicable to our method (explained in Sec. 5.3). Therefore, we propose the Dynamic Confounder Bank strategy, which is divided into two stages: 1) Use the moving average method to calculate $\mathbb{E}\boldsymbol{I}$; 2) Estimate the causal effect of $\boldsymbol{I}$ on $\boldsymbol{g}$ by mapping $\mathbb{E}\boldsymbol{I}$.

During the training phase, since the model can only process a batch of samples at a time, it cannot directly compute $\mathbb{E}\boldsymbol{I}$. To address this issue, we employ the moving average method to approximate the value of $\mathbb{E}\boldsymbol{I}$. Assume that the number of samples in a batch is N_B, and the current batch is the t-th batch,

with $N_0 = 0$ and $[\mathbb{E}\boldsymbol{I}]_0 = (0, \cdots, 0)^T$. the formula is as follows:

$$\begin{aligned} [\mathbb{E}\boldsymbol{I}]_t &= \frac{N_{t-1}}{N_{t-1}+N_B}[\mathbb{E}\boldsymbol{I}]_{t-1} + \frac{N_B}{N_{t-1}+N_B}\text{Mean}(\{\boldsymbol{I}\}_{\text{Batch}}), \\ N_t &= N_{t-1} + N_B, \end{aligned} \tag{6}$$

where $\text{Mean}(\cdot)$ denotes the mean function, $[\mathbb{E}\boldsymbol{I}]_t$ represents the estimated value of $\mathbb{E}\boldsymbol{I}$ after the first t-batches. $\{\boldsymbol{I}\}_{\text{Batch}}$ represents the set of all $\boldsymbol{I}$ values extracted from the samples within the current batch. Using Eq. (6), Our method can compute $\mathbb{E}\boldsymbol{I}$ online. Since $\boldsymbol{I}$ represents the gaze-irrelevant features, $\mathbb{E}\boldsymbol{I}$ can thus be understood as the average of all gaze-irrelevant features. When training completes, $\mathbb{E}\boldsymbol{I}$ is frozen for inference.

4.3 Implementation Details

Our method is implemented using the Pytorch framework. We employ **ResNet18** as the backbone, with the fully connected layer and global pooling layer removed from ResNet18. The batch size is set to 512. We use two identical Adam optimizers to update the parameters of FSM and CIM, respectively. The learning rate for the optimizers are set to 10^{-4}, with $\beta = (0.5, 0.95)$.

5 Experiments

5.1 Data Preprocessing

We validated our method on four commonly used gaze datasets: ETH-XGaze($\mathcal{D}_E$) [41], Gaze360($\mathcal{D}_G$) [16], MPIIFaceGaze($\mathcal{D}_M$) [43], and EyeDiap($\mathcal{D}_D$) [12]. Following previous gaze Domain Generalization studies [7, 39], we use $\mathcal{D}_E$ and $\mathcal{D}_G$ as source domain because they provide a larger gaze distribution range.

Data preparation. For $\mathcal{D}_E$, $\mathcal{D}_M$, and $\mathcal{D}_D$, we follow the technique in [43] to normalize the face image. For $\mathcal{D}_G$, we only use frontal face images and do not employ the normalization technique [43]. Above data processings are consistent with [7, 24, 36, 39]. Since $\mathcal{D}_G$ does not provide head pose labels, we generate head pose annotations for $\mathcal{D}_G$ using Mediapipe [25]. Finally, we scale all the face images from the datasets to 224×224 and normalize the pixel values to $[0, 1]$. Some of the data preprocessing codes are provided by [10].

5.2 Comparison with State-Of-The-Art Methods

For DG tasks, we compared the performance of FSCI with other SOTA gaze estimation (GE) and gaze domain generalization methods on four cross-domain tasks: $\mathcal{D}_E \rightarrow \mathcal{D}_M, \mathcal{D}_E \rightarrow \mathcal{D}_D, \mathcal{D}_G \rightarrow \mathcal{D}_M, \mathcal{D}_G \rightarrow \mathcal{D}_D$. The experimental results are shown in Tab. 1. We report the cross-domain performance of gaze estimation methods [6,8,43] according to [7]. Results demonstrate that our method not only achieves consistent improvement compared to the baseline ResNet-18 method, but the proposed FSCI framework also outperforms all SOTA GE and gaze DG methods in all cross-domain tasks. The above results prove the effectiveness of the FSCI framework.

Table 1: Comparison with state-of-the-art domain generalization methods. Results are angular error in degrees.

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
ResNet18	9.07	10.15	9.75	11.41
Full-Face [43]	12.35	30.15	11.13	14.42
Dilated-Net [6]	-	-	18.45	23.88
CA-NET [8]	-	-	27.13	31.41
PureGaze [7]	7.08	7.48	9.28	9.32
Xu *et al.* [39]	6.50	7.44	7.55	9.03
FSCI(ours)	**5.79**	**6.96**	**7.06**	**7.99**

5.3 Ablation Study

To prove the effectiveness of each component of the FSCI framework, we conducted two ablation experiments. Tab. 2 studies the impact of the backbone network, FSM and CIM. Tab. 3 demonstrates the effectiveness of the Dynamic Confounder Bank by comparing it to other processing strategies that alleviate the impact of factors causing spurious correlations in SOTA casual intervention methods. For brevity, in all of the following content, we refer to the factors causing spurious correlations as *confounding factors*.

Ablation Study on Backbone Architecture and Proposed Modules. Results in Tab. 2 demonstrate three conclusions: (1) Both FSCI-ResNet18 and FSCI-ResNet50 show superior performance in four cross-domain experiments compared to ResNet18 and ResNet50, indicating that the FSCI framework is applicable to different backbones. While ResNet50 generally preforms poorly on $\mathcal{D}_G \rightarrow \mathcal{D}_D$, as noted in [7, 41], FSCI-ResNet50 still improves by 13.8% over ResNet50. (2) In **FSCI-ResNet18×3**, we replace the FSM with three ResNet18 trained by $\mathcal{L}_{sep}$ for feature separation. Compared to the FSCI, cross-domain performance of **FSCI-ResNet18×3** drops in all tasks, proves the effectiveness of the FSM. (3) Without the CIM , the cross-domain accuracy of the FSCI framework drops in seven out of eight settings, demonstrating the effectiveness of the proposed CIM . For detailed experimental results on the FSM and the separated features, please refer to the supplementary materials.

Ablation Study on Processing Strategies for Confounding Factors. In this part, we analyze the effectiveness of the Dynamic Confounder Bank, the key component of implementing causal intervention in FSCI framework. In causal intervention for computer vision methods, a common approach to addressing confounding factors is to construct a confounder dictionary [23, 35, 40]. This dictionary stores the feature vectors corresponding to discrete confounding factors. The construction of a confounder dictionary can be divided into two steps: 1) Assigning confounding type labels to each sample in the given dataset; 2) Extracting features of all samples in the dataset using a pre-trained feature

Table 2: Ablation study on backbone network and proposed modules. Method 4, 8 are our proposed methods with different backbones. In ResNet×3, we replace the FSM with 3 ResNet trained by **FSCI-ResNet×3** for feature seperation. Results are angular error in degrees.

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
1 ResNet18	9.07	10.15	9.75	11.41
2 FSCI-ResNet18×3	8.72	8.01	8.24	9.18
3 FSCI-ResNet18 w/o CIM	7.67	7.48	8.33	9.75
4 FSCI-ResNet18	**5.79**	**6.96**	**7.06**	**7.99**
5 ResNet50	7.59	8.70	8.75	11.83
6 FSCI-ResNet50×3	5.96	6.58	7.17	12.99
7 FSCI-ResNet50 w/o CIM	6.17	7.23	8.07	**10.00**
8 FSCI-ResNet50	**5.47**	**6.68**	**6.19**	10.20

Table 3: Ablation study on different processing strategies of gaze-irrelevant feature. The experiment compared the proposed Dynamic Confounder Bank strategy with the confounder dictionary strategy. Results are angular error in degrees.

Task	Confounder dictionaries			FSCI
	Id	Id×Skin	Id×Skin×Illum	
$\mathcal{D}_E \rightarrow \mathcal{D}_M$	5.96	**5.77**	5.83	5.79
$\mathcal{D}_E \rightarrow \mathcal{D}_D$	8.38	8.18	8.25	**6.96**

extractor and categorizing them according to each sample's corresponding confounding type label, then using the class mean of each confounding label as the confounding variable corresponding to that label. In Tab. 3, we constructed three confounder dictionaries according to three potential confounding factors in the Gaze Estimation task, *i.e.* Identity, complexion, and illumination:

- **Id:** Using the identity of individuals in each sample image as the confounding label (15 confounding variables).
- **Id×Skin:** Using the Cartesian product of the identity and complexion as the confounding label (30 confounding variables).
- **Id×Skin×Illum:** Using the Cartesian product of the identity, complexion, and image brightness (divided into three brightness intervals) as the confounding label (90 confounding variables).

As shown in Tab. 3, there is no clear superiority between the two strategies for addressing confounding factors in the $\mathcal{D}_E \rightarrow \mathcal{D}_M$ cross-domain task. However, in the $\mathcal{D}_E \rightarrow \mathcal{D}_D$ cross-domain task, the Dynamic Confounder Bank strategy significantly outperforms the confounder dictionary strategy. Compared to the Dynamic Confounder Bank strategy, building a confounder dictionary requires additional sample labels, which incurs a certain cost of manual annotation. Moreover, the confounder dictionary stores confounding factors in a discrete form. Since it is impossible to identify all confounding factors in the dataset, the

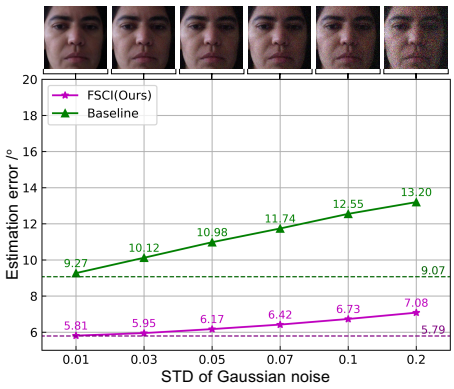

Fig. 4: Estimation errors under different Gaussian noise levels in $\mathcal{D}_E \rightarrow \mathcal{D}_M$. Dashed lines represent the estimation errors without noise.

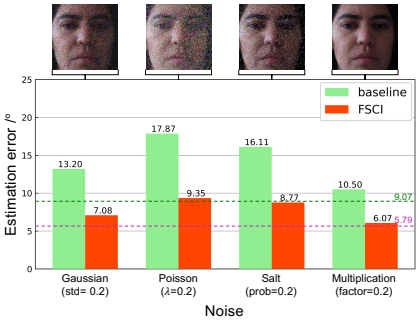

Fig. 5: Gaze estimation errors under different types of noise in $\mathcal{D}_E \rightarrow \mathcal{D}_M$. The content inside the brackets in the x-axis labels represents the parameters of the corresponding noise. Dashed lines represent the estimation errors without noise.

construction of a confounder dictionary inevitably faces the issue of missing confounding factors. While solving the above issues, the Dynamic Confounder Bank strategy also achieves relatively better performance in cross-domain tasks.

5.4 Verification of Causal Intervention

Theoretically, the spurious correlation $\boldsymbol{I} \dashrightarrow \boldsymbol{G}$ and $\boldsymbol{H} \dashrightarrow \boldsymbol{G}$ will be both weakened if the FSCI framework successfully alleviates the influence of $\boldsymbol{I}$, as explained in Sec. 3 and Fig. 2. Following this theory, we conduct further analysis to verify the effect of the FSCI framework in the following sections.

Impact of Spurious Correlation $\boldsymbol{I} \dashrightarrow \boldsymbol{G}$. To quantificat the impact of $\boldsymbol{I} \dashrightarrow \boldsymbol{G}$, we aim to examine the model's resilience to disturbance of gaze-irrelevant factors. However, it is difficult to disturb gaze-irrelevant factors such as identity while keeping gaze and head pose information completely the same. Alternatively, we apply random noises to the input image in the $\mathcal{D}_E \rightarrow \mathcal{D}_M$ task as an approximation. As shown in Fig. 4, the proposed FSCI framework demonstrates significantly better resistance against Gaussian Noise than the Baseline model. When applying a Gaussian Noise with 0.2 STD, the estimation error of the baseline model increases for 4.13°. On the contrary, the estimation error of the FSCI framework only increases for 1.29°, which is 68.8% smaller than the baseline. In Fig. 5, we apply 3 more different types of noise, the FSCI framework consistently shows better stability. Above results prove that the proposed FSCI framework weakens the spurious correlation $\boldsymbol{I} \dashrightarrow \boldsymbol{G}$ significantly during gaze estimation, which improves the generalization ability of the model.

Impact of Spurious Correlation $\boldsymbol{H} \dashrightarrow \boldsymbol{G}$. To verify the spurious correlation $\boldsymbol{H} \dashrightarrow \boldsymbol{G}$ learned by the gaze estimation model, we visualize the distributions

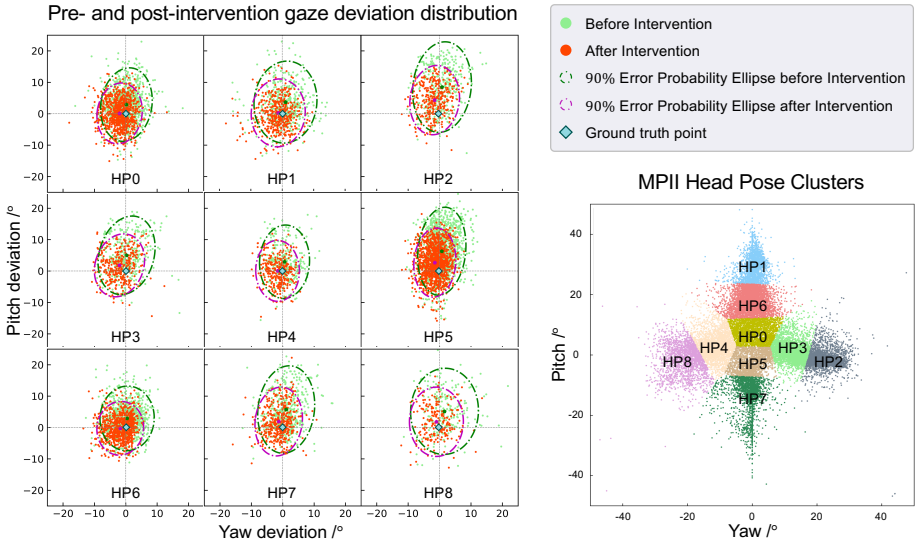

Fig. 6: The distribution of the gaze estimation deviation within different head pose clusters before and after causal intervention in $\mathcal{D}_E \rightarrow \mathcal{D}_M$. Different categories of head poses were obtained by K-Means clustering.

of gaze estimation deviation before and after causal intervention in different head pose clusters in the $\mathcal{D}_E \rightarrow \mathcal{D}_M$ task, as shown in Fig. 6. The estimation results before intervention are gaze predictions from the FSM. Ideally, without the spurious correlation $\boldsymbol{H} \dashrightarrow \boldsymbol{G}$, the distribution of the deviation should be identical across different head pose clusters, *i.e.* distributes in a circle around the $(0, 0)$ point. Obviously, the distributions of deviation before intervention drift from the $(0, 0)$ point. The spurious correlation $\boldsymbol{H} \dashrightarrow \boldsymbol{G}$ seems stronger in head pose cluster 2 (HP2) since the distribution of deviation drifts further. After we apply causal intervention by CIM, the distributions of estimation deviation become more compact, and the centers of the distributions are closer to $(0, 0)$ point. These results indicate that the CIM successfully reduce the impact of spurious correlation $\boldsymbol{H} \dashrightarrow \boldsymbol{G}$. Consequently, the FSCI framework demonstrate better cross-domain accuracy after causal intervention.

Fundamentally, the source of the spurious correlation $\boldsymbol{H} \dashrightarrow \boldsymbol{G}$ is that gaze-irrelevant features $\boldsymbol{I}$ is the common cause of head pose feature $\boldsymbol{H}$ and gaze feature $\boldsymbol{G}$. Thus, we further verify the gaze estimation error with respect to $\boldsymbol{H}$ and $\boldsymbol{I}$. As shown in Fig. 7, we visualize the heatmaps of gaze estimation errors under different head poses for various complexions before and after casual intervention. Due to clear complexion categorization, we select it as a representative factor of $\boldsymbol{I}$. Ideally, if $\boldsymbol{G}$ is completely free from the influence of $\boldsymbol{I}$ and $\boldsymbol{H}$, the distribution of errors with respect to head pose should be the same across different complexions, *i.e.*, the distribution of Fig. 7 (a-1) and (b-1) should be the same, and Fig. 7 (c-1), (c-2) should be zeros. Obviously, values of Fig. 7 (c-2) is smaller than (c-1), indicates that the differences of estimation errors

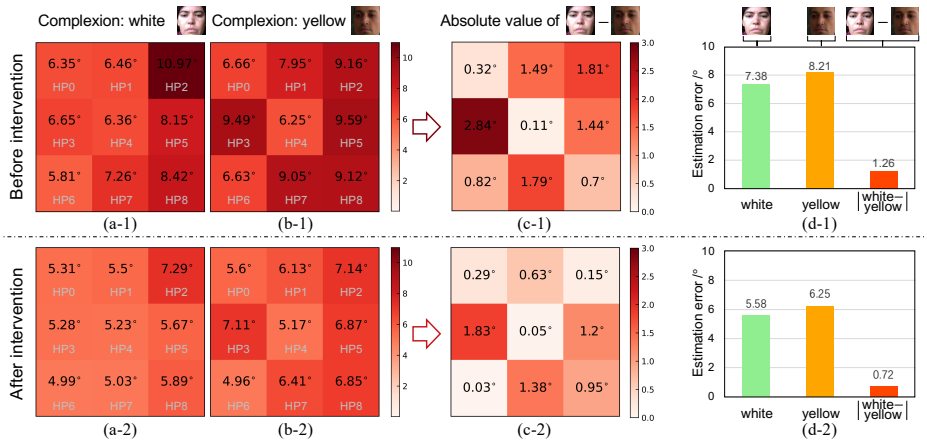

Fig. 7: (a), (b) Heatmaps of gaze estimation errors under different head poses for various complexions. (c) Absolute differences between (a) and (b). (d) Average estimation errors on samples from different complexions. After intervention, the differences in estimation errors between different complexions and head poses are reduced.

between different complexions are reduced after intervention. In Fig. 7, (d), we calculate the average estimation error of different complexions. The differences of estimation error between complexions is reduced by 42.9% after intervention. This phenomenon confirms our theory that the FSCI framework alleviates the impact of the spurious correlation $\boldsymbol{H} \dashrightarrow \boldsymbol{G}$ by minimizing the influence of gaze-irrelevant features $\boldsymbol{I}$. Note that the FSCI framework only weakens the spurious correlation $\boldsymbol{H} \dashrightarrow \boldsymbol{G}$. There is still a true correlation between head pose and gaze $\boldsymbol{H} \rightarrow \boldsymbol{G}$. Thus, it is normal that there is a difference of estimation errors between various head pose clusters. An intuitive explanation is that the difficulty of gaze estimation task varies under different head poses.

6 Limitations and Conclusions

Limitation. The proposed FSCI framework effectively mitigates the impact of confounding factors. However, the performance of CIM is somewhat dependent on the quality of the separated features. Although our experiments demonstrate that FSM is effective and enables CIM to work well, we believe that better feature separation would further enhance our final results.

Conclusion. In this paper, we propose a gaze domain generalization method based on causal inference, named the FSCI framework. To alleviate the influence of gaze-irrelevant factors during gaze estimation, the proposed FSCI framework separates different features and employs causal intervention to the gaze-irrelevant features through the Dynamic Confounder Bank strategy. Experimental results show that the FSCI framework outperforms SOTA gaze DG methods in various cross-domain tasks. Further analysis demonstrates that the FSCI framework successfully reduce the impact of spurious correlation $\boldsymbol{I} \dashrightarrow \boldsymbol{G}$.

References

1. Alletto, S., Palazzi, A., Solera, F., Calderara, S., Cucchiara, R.: Dr(eye)ve: A dataset for attention-based tasks with applications to autonomous and assisted driving. In: CVPRW (June 2016)
2. Bao, Y., Cheng, Y., Liu, Y., Lu, F.: Adaptive feature fusion network for gaze tracking in mobile tablets. In: ICPR. pp. 9936–9943. IEEE (2021)
3. Bao, Y., Liu, Y., Wang, H., Lu, F.: Generalizing gaze estimation with rotation consistency. In: CVPR. pp. 4207–4216 (2022)
4. Castner, N., Kuebler, T.C., Scheiter, K., Richter, J., Eder, T., Huettig, F., Keutel, C., Kasneci, E.: Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing. In: ETRA. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3379155.3391320, https://doi.org/10.1145/3379155.3391320
5. Chen, L., Zheng, Y., Niu, Y., Zhang, H., Xiao, J.: Counterfactual samples synthesizing and training for robust visual question answering. PAMI (2023)
6. Chen, Z., Shi, B.E.: Appearance-based gaze estimation using dilated-convolutions. In: ACCV. pp. 309–324. Springer (2018)
7. Cheng, Y., Bao, Y., Lu, F.: Puregaze: Purifying gaze feature for generalizable gaze estimation. In: AAAI. vol. 36, pp. 436–443 (2022)
8. Cheng, Y., Huang, S., Wang, F., Qian, C., Lu, F.: A coarse-to-fine adaptive network for appearance-based gaze estimation. In: AAAI. vol. 34, pp. 10623–10630 (2020)
9. Cheng, Y., Lu, F., Zhang, X.: Appearance-based gaze estimation via evaluation-guided asymmetric regression. In: ECCV. pp. 100–115 (2018)
10. Cheng, Y., Wang, H., Bao, Y., Lu, F.: Appearance-based gaze estimation with deep learning: A review and benchmark. arXiv preprint arXiv:2104.12668 (2021)
11. Cheng, Y., Zhang, X., Lu, F., Sato, Y.: Gaze estimation by exploring two-eye asymmetry. TIP **29**, 5259–5272 (2020)
12. Funes Mora, K.A., Monay, F., Odobez, J.M.: Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In: ETRA. pp. 255–258 (2014)
13. Guo, Z., Yuan, Z., Zhang, C., Chi, W., Ling, Y., Zhang, S.: Domain adaptation gaze estimation by embedding with prediction consistency. In: ACCV (2020)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
15. Katsini, C., Abdrabou, Y., Raptis, G.E., Khamis, M., Alt, F.: The role of eye gaze in security and privacy applications: Survey and future hci research directions. In: CHI. pp. 1–21 (2020)
16. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: ICCV. pp. 6912–6921 (2019)
17. Kerr-Gaffney, J., Harrison, A., Tchanturia, K.: Eye-tracking research in eating disorders: A systematic review. International Journal of Eating Disorders **52**(1), 3–27 (2019). https://doi.org/https://doi.org/10.1002/eat.22998, https://onlinelibrary.wiley.com/doi/abs/10.1002/eat.22998
18. Konrad, R., Angelopoulos, A., Wetzstein, G.: Gaze-contingent ocular parallax rendering for virtual reality. TOG **39**(2) (jan 2020). https://doi.org/10.1145/3361330, https://doi.org/10.1145/3361330
19. Kothari, R., De Mello, S., Iqbal, U., Byeon, W., Park, S., Kautz, J.: Weakly-supervised physically unconstrained gaze estimation. In: CVPR. pp. 9980–9989 (2021)

20. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: CVPR. pp. 2176–2184 (2016)
21. Lian, D., Hu, L., Luo, W., Xu, Y., Duan, L., Yu, J., Gao, S.: Multiview multitask gaze estimation with deep convolutional neural networks. IEEE transactions on neural networks and learning systems **30**(10), 3010–3023 (2018)
22. Liu, C., Chen, Y., Tai, L., Ye, H., Liu, M., Shi, B.E.: A gaze model improves autonomous driving. In: ETRA. pp. 1–5 (2019)
23. Liu, Y., Li, G., Lin, L.: Cross-modal causal relational reasoning for event-level visual question answering. PAMI (2023)
24. Liu, Y., Liu, R., Wang, H., Lu, F.: Generalizing gaze estimation with outlier-guided collaborative adaptation. In: ICCV. pp. 3835–3844 (2021)
25. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
26. Park, S., Mello, S.D., Molchanov, P., Iqbal, U., Hilliges, O., Kautz, J.: Few-shot adaptive gaze estimation. In: ICCV. pp. 9368–9377 (2019)
27. Pearl, J., Glymour, M., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)
28. Pearl, J., et al.: Models, reasoning and inference. Cambridge, UK: CambridgeUniversityPress **19**(2), 3 (2000)
29. Rahal, R.M., Fiedler, S.: Understanding cognitive and affective mechanisms in social psychology through eye-tracking. Journal of Experimental Social Psychology **85**, 103842 (2019)
30. Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., Wetzstein, G.: Saliency in vr: How do people explore virtual environments? TVCG **24**(4), 1633–1642 (2018). https://doi.org/10.1109/TVCG.2018.2793599
31. Tawari, A., Chen, K.H., Trivedi, M.M.: Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation. In: ITSC. pp. 988–994. IEEE (2014)
32. Terzioğlu, Y., Mutlu, B., Şahin, E.: Designing social cues for collaborative robots: the role of gaze and breathing in human-robot collaboration. In: HRI. pp. 343–357 (2020)
33. Wang, H., Dong, X., Chen, Z., Shi, B.E.: Hybrid gaze/eeg brain computer interface for robot arm control on a pick and place task. In: EMBC. pp. 1476–1479. IEEE (2015)
34. Wang, K., Zhao, R., Su, H., Ji, Q.: Generalizing eye tracking with bayesian adversarial learning. In: CVPR. pp. 11907–11916 (2019)
35. Wang, T., Huang, J., Zhang, H., Sun, Q.: Visual commonsense r-cnn. In: CVPR. pp. 10760–10770 (2020)
36. Wang, Y., Jiang, Y., Li, J., Ni, B., Dai, W., Li, C., Xiong, H., Li, T.: Contrastive regression for domain adaptation on gaze estimation. In: CVPR. pp. 19376–19385 (2022)
37. Wang, Z., Zhao, Y., Lu, F.: Control with vergence eye movement in augmented reality see-through vision. In: VRW. pp. 548–549. IEEE (2022)
38. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Bach, F., Blei, D. (eds.) ICML. vol. 37, pp. 2048–2057. PMLR, Lille, France (07–09 Jul 2015), https://proceedings.mlr.press/v37/xuc15.html
39. Xu, M., Wang, H., Lu, F.: Learning a generalized gaze estimator from gaze-consistent feature. In: AAAI. vol. 37, pp. 3027–3035 (2023)

40. Yang, D., Chen, Z., Wang, Y., Wang, S., Li, M., Liu, S., Zhao, X., Huang, S., Dong, Z., Zhai, P., et al.: Context de-confounded emotion recognition. In: CVPR. pp. 19005–19015 (2023)
41. Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O.: Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In: ECCV. pp. 365–381. Springer (2020)
42. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: CVPR. pp. 4511–4520 (2015)
43. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: CVPRW. pp. 51–60 (2017)