Gaze from Origin: Learning for Generalized Gaze Estimation by Embedding the Gaze Frontalization Process

Mingjie Xu[1], **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
{xumingjie, lufeng}@buaa.edu.cn

Abstract

Gaze estimation aims to accurately estimate the direction or position at which a person is looking. With the development of deep learning techniques, a number of gaze estimation methods have been proposed and achieved state-of-the-art performance. However, these methods are limited to within-dataset settings, whose performance drops when tested on unseen datasets. We argue that this is caused by infinite and continuous gaze labels. To alleviate this problem, we propose using gaze frontalization as an auxiliary task to constrain gaze estimation. Based on this, we propose a novel gaze domain generalization framework named Gaze Frontalization-based Auxiliary Learning (GFAL) Framework which embeds the gaze frontalization process, i.e., guiding the feature so that the eyeball can rotate and look at the front (camera), without any target domain information during training. Experimental results show that our proposed framework is able to achieve state-of-the-art performance on gaze domain generalization task, which is competitive with or even superior to the SOTA gaze unsupervised domain adaptation methods.

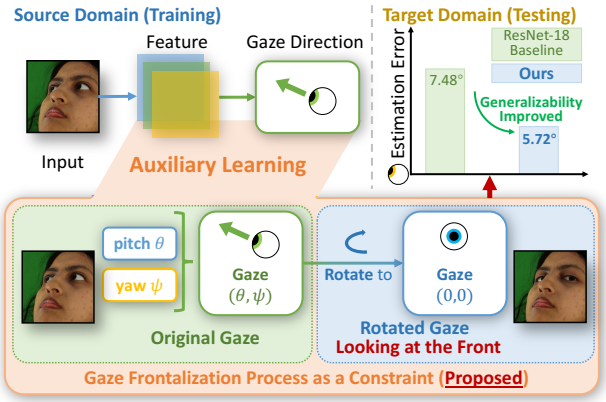

Figure 1: Overall idea of our proposed gaze domain generalization framework GFAL, which is able to improve cross-domain gaze estimation performance.

Introduction

Gaze information is important for real applications. It indicates the direction or position at which a person is looking, and is widely used in many scenarios, such as augmented reality (Wang, Zhao, and Lu 2022) and autonomous driving (Mole et al. 2021). To obtain this information, a number of gaze estimation methods have been proposed. In the early years, model-based gaze estimation methods were popular in this field. Although they were able to accurately estimate gaze, they required dedicated devices and calibration (Sun, Liu, and Sun 2015). To tackle this problem, appearance-based gaze estimation approaches have been widely proposed in recent years (Lu et al. 2011, 2014), especially deep-learning-based gaze estimation methods (Zhang et al. 2017a; Chen and Shi 2018). Such methods only use RGB images as input and are capable of achieving competitive performance.

Although all of these gaze estimation methods are capable of achieving state-of-the-art performance when trained and tested on the same dataset (*i.e.*the source domain), they often suffer performance degradation on unseen test datasets (*i.e.*the target domain). This problem is defined as cross-domain gaze estimation, which has not been fully addressed and remains difficult (Cheng, Bao, and Lu 2022; Wang et al. 2022; Bao et al. 2022).

The reason behind the difficulty is that gaze labels are infinite and continuous. Typical classification or recognition tasks output finite and discrete labels. However, gaze estimation is a regression task and regresses continuous Euler angles of gaze directions, which form an infinite set of labels. Consequently, accurately estimating these gaze labels is challenging, particularly in cross-domain scenarios. Moreover, infinite and continuous gaze labels is one of the major causes for overfitting in the finite training dataset.

We propose using an auxiliary task for auxiliary learning to constrain gaze estimation, which can help to alleviate these problems. We argue that gaze frontalization is a proper choice for such an auxiliary task because: 1) Gaze frontalization and gaze estimation are homogeneous. 2) Gaze frontalization is also easier than gaze estimation, which can help constrain gaze estimation by performing multi-task learning or serving as a regularization term.

Based on these analyses, we propose a novel gaze domain generalization framework named *Gaze Frontalization-based*

*Corresponding Author.

Auxiliary Learning (GFAL) Framework, which can improve cross-domain gaze estimation performance without any target domain data or label for training. First, we extract the features from a single face image. Second, we embed the gaze frontalization process to guide the features so that the eyeball can rotate and look at the front (camera), as shown in Fig. 1. Our main contributions are as follows:

- We systematically analyzed the cross-domain regression problem of gaze estimation and proposed using a proper auxiliary task, *i.e*., the gaze frontalization task, to tackle this problem.
- Based on this, we propose a novel gaze domain generalization framework named GFAL, by embedding the gaze frontalization process, which aims to guide the features so that the eyeball can rotate and look at the front (camera). This framework is flexible, as it can be used with image warping or multiple gaze redirection methods to improve cross-domain gaze estimation performance.
- The experimental results show that GFAL framework can achieve significant performance improvements of 23.53%, 11.32%, 6.75% and 9.11% over the baseline on ETH-to-MPII, ETH-to-Diap, Gaze360-to-MPII and Gaze360-to-Diap tasks, respectively, and surpass all SOTA gaze domain generalization methods.

Related Works

Gaze Estimation. Recent appearance-based gaze estimation methods usually use deep learning techniques, such as Convolution Neural Network (CNN) (Zhang et al. 2017b,a; Cheng, Lu, and Zhang 2018; Cheng et al. 2020b; Fischer, Chang, and Demiris 2018; Park, Spurr, and Hilliges 2018; Chen and Shi 2018; Yu, Liu, and Odobez 2019a; Cheng et al. 2020a) or Vision Transformer (Dosovitskiy et al. 2021; Cheng and Lu 2022; O Oh, Chang, and Choi 2022). Thanks to the effective architecture and various datasets (Zhang et al. 2020; Kellnhofer et al. 2019; Zhang et al. 2017b; Funes Mora, Monay, and Odobez 2014), this kind of gaze estimation methods can achieve excellent performance. For example, Park *et al*.proposed a method with the help of facial landmark extraction (Park et al. 2018). Bao *et al*.proposed a lightweight gaze estimation network designed for mobile devices, combining information on the face, two eyes, and eye positions (Bao et al. 2021). Although these gaze estimation approaches can all achieve SOTA performance, they often suffer from poor gaze estimation accuracy when tested on unseen target domain. Addressing such cross-domain gaze estimation problem is necessary for further application.

To tackle the abovementioned cross-domain gaze estimation problem, various approaches have been proposed, including Supervised Domain Adaptation (SDA) methods, Unsupervised Domain Adaptation (UDA) methods, and Domain Generalization (DG) methods.

Gaze SDA (Krafka et al. 2016; He et al. 2019; Yu, Liu, and Odobez 2019b) and UDA (Kellnhofer et al. 2019; Lee et al. 2022) methods usually fine-tune the source domain pretrained model on a few labeled or unlabeled target domain samples to improve cross-domain gaze estimation performance, respectively. For example, gaze SDA methods are often based on meta-learning (Park et al. 2019), gaze difference (Liu et al. 2018, 2019) and gaze decomposition (Chen and Shi 2020, 2022), while gaze UDA methods are often based on adversarial learning (Wang et al. 2019; Lahiri, Agarwalla, and Biswas 2018), teacher-student networks (He et al. 2019; Liu et al. 2021), representation learning (Guo et al. 2021; Wang et al. 2022), rotation consistency (Bao et al. 2022) and jitter (Liu et al. 2022). Although these approaches can achieve excellent results in cross-domain gaze estimation, samples or labels on target domain are difficult to collect in real world, which limits their applications.

To tackle the data collection problem in real world, DG methods are proposed for cross-domain gaze estimation, which do not require any target domain information when training. These methods typically use adversarial learning (Cheng, Bao, and Lu 2022) or adversarial attack (Xu, Wang, and Lu 2023) to eliminate or disturb gaze-irrelevant factors or features. Moreover, some gaze SDA, UDA or redirection methods also contain gaze DG modules, such as (Park et al. 2019; Bao et al. 2022; Wang et al. 2022; Lee et al. 2022). However, their cross-domain performances still have room for improvement.

Many of these gaze SDA, UDA or DG methods use auxiliary tasks. Different from our proposed GFAL, the auxiliary tasks used by these methods cannot solve the infinite and continuous gaze problem.

Gaze Redirection. Gaze redirection methods aim to generate face or eye images looking at the given target direction. Recently, many gaze redirection methods have been proposed, such as DeepWarp (Ganin et al. 2016), ST-ED (Zheng et al. 2020), CUDA-GHR (Jindal and Wang 2023) and (Yu, Liu, and Odobez 2019b). These methods are usually used to extend the training dataset for gaze estimation (Zheng et al. 2020; Jindal and Wang 2023). However, gaze estimation models still suffer from the problems stated above, which cannot improve cross-domain gaze estimation performance. Moreover, some methods try to use gaze redirection to perform gaze representation learning for better calibration, such as (Park et al. 2019; Yu, Liu, and Odobez 2019b; Yu and Odobez 2020), but they all rely on labeled calibration samples, which are hard to obtain in the real world.

Motivation and Key Idea

Formulation

Gaze Estimation. Formally, we first define the gaze estimation task, as discussed in (Zhang et al. 2020). Given the input face image $\mathbf{x}$, the gaze direction $\mathbf{g}$ can be estimated via:

$$\mathbf{z} = \mathcal{F}(\mathbf{x}), \mathbf{g} = \mathcal{G}(\mathbf{z}), \tag{1}$$

where $\mathcal{F}$ is a feature extractor, $\mathcal{G}$ is a fully connected (FC) layer and $\mathbf{z}$ is the extracted feature. Note that the gaze direction $\mathbf{g}$ is expressed by the Euler angle (θ, ψ) defined by pitch θ and yaw ψ.

Cross-Domain Gaze Estimation. Typical gaze estimation models are often trained and tested on the source domain $\mathcal{D}_s$ and achieve excellent performance. However, if these models are tested on an unseen target domain $\mathcal{D}_t$, their performance usually degrades. This problem is called cross-domain gaze estimation.

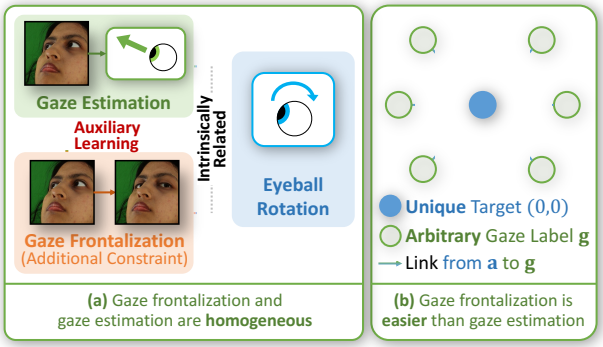

Figure 2: Illustration of our motivation and solution.

Problem Analysis

The cross-domain gaze estimation problem is difficult to tackle, because gaze labels are infinite and continuous, which is one of the major causes for overfitting in the finite training dataset. Gaze estimation is a kind of regression task. As described above, the gaze estimation model outputs pitch θ and yaw ψ. Arbitrary continuous values can be used to express θ and ψ. Therefore, there are infinite (θ, ψ)-s, forming infinite and continuous gaze labels. This characteristic poses challenges in accurately learning the mapping from images to gaze, especially in cross-domain scenarios. This is more difficult than classification or recognition tasks because the labels of the latter tasks are finite and discrete.

Why We Use Gaze Frontalization as an Auxiliary Task?

Based on the above analyses, we propose the use of an auxiliary task to constrain gaze estimation to tackle the cross-domain gaze estimation problem. This auxiliary task should help to solve the abovementioned problem. We use gaze frontalization as an auxiliary task, which involves rotating the eyeballs to make them look at the front (camera), *i.e.*, the $(0, 0)$ direction. We argue that the gaze frontalization task is a proper choice for such an auxiliary task for the following reasons: 1) Gaze frontalization and gaze estimation are homogeneous, involving the utilization of face images as input and the association with eyeball rotation. 2) Gaze frontalization is also easier than gaze estimation, as the target of gaze frontalization corresponds to a unique direction (0,0), while gaze estimation takes arbitrary gazes as target. Thus, Gaze frontalization (easier) can help constrain gaze estimation (harder) by performing multi-task learning or serving as a regularization term. The frontal gaze (0,0) can serve as anchors to constrain gaze directions and eyeball rotation.

Following the above, our motivation is to find out how to use gaze frontalization to constrain gaze estimation. To achieve this, we embed the gaze frontalization process to guide the extracted gaze features so that the eyeball can rotate to (0,0), further improving cross-domain gaze estimation performance. Note that it is not helpful to using gazes other than $(0, 0)$ to constrain gaze estimation because arbitrary gazes pose greater difficulty.

GFAL Framework

Based on the above analyses, we propose a novel gaze domain generalization framework named *Gaze Frontalization-based Auxiliary Learning (GFAL) Framework*, based on an embedded gaze frontalization process, *i.e.*, forcing the gaze feature $\mathbf{z}$ to rotate the eyeballs so that the eyeballs can look at $(0, 0)$, for auxiliary learning. GFAL framework includes 3 parts: 1) Gaze Estimation Network, 2) Gaze Frontalization Module and 3) Consistency Loss, as shown in Fig. 3(a).

Gaze Estimation Network

First, we propose a Gaze Estimation Network for gaze and head orientation estimation, as shown in Fig. 3(b).

Gaze Estimation. As described above, gaze estimation uses a feature extractor $\mathcal{F}$ and a FC layer $\mathcal{G}$, takes a face image $\mathbf{x}$ as input and outputs the gaze direction $\hat{\mathbf{g}} = \mathcal{G}(\mathcal{F}(\mathbf{x}))$, which is defined by pitch θ and yaw ψ.

We use the commonly used L_1 distance between the estimated gaze direction $\hat{\mathbf{g}}$ and the ground truth gaze direction $\mathbf{g}$ for training:

$$\mathcal{L}_{\mathcal{G}} = \|\hat{\mathbf{g}} - \mathbf{g}\|_1. \tag{2}$$

Head Orientation Estimation. If the training dataset provides head orientation labels, we also estimate head orientations $\hat{\mathbf{h}} = \mathcal{H}(\mathcal{F}(\mathbf{x}))$ using the feature extractor $\mathcal{F}$ and a FC layer $\mathcal{H}$ whose architecture is the same as that of $\mathcal{G}$, since this information is widely used in and helpful for gaze estimation methods (Park et al. 2019; Wang et al. 2019):

$$\mathcal{L}_{\mathcal{H}} = \|\hat{\mathbf{h}} - \mathbf{h}\|_1, \tag{3}$$

where $\hat{\mathbf{h}}$ is the estimated head orientation and $\mathbf{h}$ is the ground truth head orientation label.

Gaze Frontalization Module

Goal and Procedure. The goal of this module is to embed the gaze frontalization process, *i.e.*, forcing the feature $\mathbf{z}$ to rotate the eyeballs in $\mathbf{x}$ so that they look at $(0, 0)$ while keeping the head orientation unchanged, for auxiliary learning.

To achieve this goal, we feed $\mathbf{z}$ and $\mathbf{x}$ into this module to obtain the output face image $\mathbf{x}_{fro}$, in which $\mathbf{z}$ is able to make the eyeball in $\mathbf{x}$ rotate to $(0, 0)$, *i.e.*, looking at the front (camera), from the direction (θ, ψ). Then, we optimize the loss function $\mathcal{L}_{fro}$ of this module.

This procedure can be implemented using various strategies, such as image warping, gaze redirection, and 3D reconstruction of face images.

Implementation. Here, we design a strategy based on image warping, as shown in Fig. 3(c). We first feed the extracted feature $\mathbf{z}$ into a warping field generator $\mathcal{W}$ to obtain the warping field $\mathbf{M}$:

$$\mathbf{M} = \mathcal{W}(\mathbf{z}). \tag{4}$$

For every position (x_{rot}, y_{rot}) in the output image $\mathbf{x}_{fro}$, there is a corresponding position (x, y) in the input image $\mathbf{x}$, which can be obtained from the warping field $\mathbf{M}$. $\mathcal{W}$ is implemented using a ResNet-like decoder (Cheng, Bao, and Lu 2022). Note that the size of $\mathbf{M}$ is the same as that of $\mathbf{x}$.

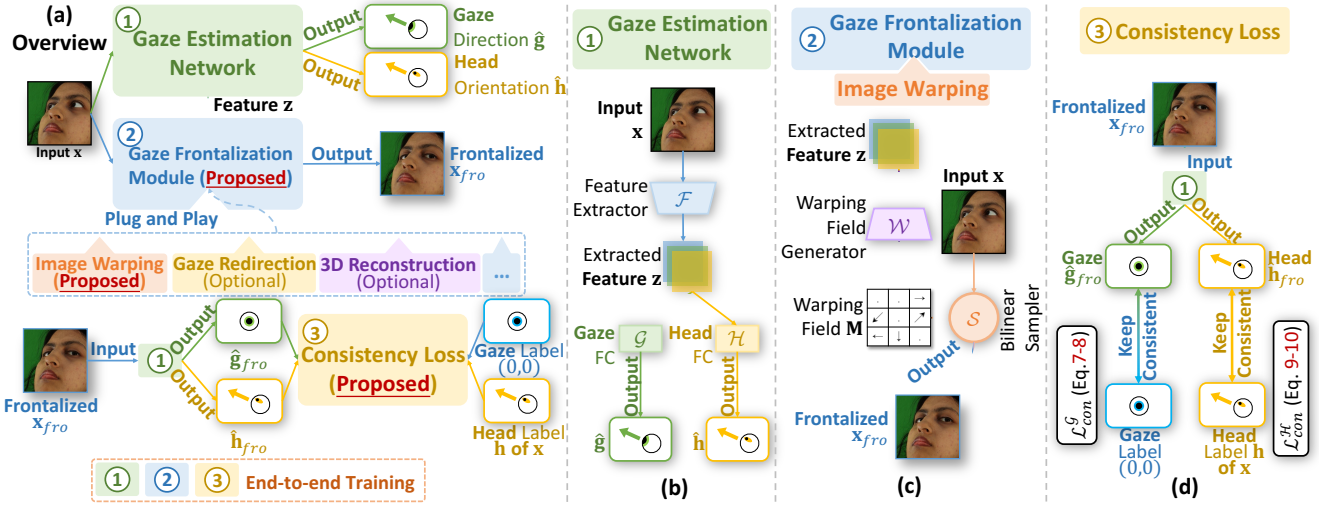

Figure 3: Our proposed gaze domain generalization framework.

Then we apply the output warping field $\mathbf{M}$ to the input face image $\mathbf{x}$ to obtain the output warped face image $\mathbf{x}_{fro}$:

$$\mathbf{x}_{fro} = \mathcal{S}(\mathbf{x}, \mathbf{M}), \tag{5}$$

where $\mathcal{S}$ is a bilinear sampler (Jaderberg et al. 2015) used to process bilinear interpolation as the output warping field $\mathbf{M}$ needs to be integer-valued but is real-valued.

To ensure that the eyes in $\mathbf{x}_{fro}$ indeed look at $(0, 0)$ with the head orientations unchanged, we introduce a reference face image $\mathbf{x}_{ref}$, which has eyeballs looking at $(0, 0)$ and the same head orientation and identity as $\mathbf{x}$, to constrain the generation of $\mathbf{x}_{fro}$. We maximize Multi-Scale Structural Similarity (MS-SSIM) (Wang, Simoncelli, and Bovik 2003) between $\mathbf{x}_{ref}$ and $\mathbf{x}_{fro}$:

$$\mathcal{L}_{fro} = 1 - \text{MS-SSIM}(\mathbf{x}_{ref}, \mathbf{x}_{fro}). \tag{6}$$

Note that there are various choices for $\mathcal{L}_{fro}$, such as the L_1, L_2 distance and Structural Similarity (SSIM) (Wang et al. 2004). The comparison of different choices of $\mathcal{L}_{fro}$ is shown in Tab. 4, in which MS-SSIM achieves the best performance.

The selection of $\mathbf{x}_{ref}$ is described in Algorithm 1. Given an input face image $\mathbf{x}$, we first select the face images $\mathcal{D}_{ide}$ that share the same identity as $\mathbf{x}$ to form the candidate set of reference images $\mathbf{x}_{ref}$ of $\mathbf{x}$. Then, we iterate over all the images $\mathbf{x}_k$ in $\mathcal{D}_{ide}$ to find the $\mathbf{x}_K$ with the gaze direction closest to $(0, 0)$ and the head orientation most similar to the head orientation of $\mathbf{x}$. Finally, $\mathbf{x}_K$ is the expected $\mathbf{x}_{ref}$.

Note that the Gaze360 dataset (Kellnhofer et al. 2019) does not provide head orientation labels and reliable identity labels. To address this problem, we use SSIM (Wang et al. 2004) to measure the head orientation similarity instead and use CosFace (Wang et al. 2018) and k-means (MacQueen 1967) to generate identity labels. Specifically, we use $1 - \text{SSIM}(\mathbf{x}_k, \mathbf{x})$ to replace $\text{Angular}(\mathbf{h}_k, \mathbf{h})$ in Line 8 of Algo. 1 and use $\text{Angular}(\mathbf{g}_k, (0, 0))/180$ to replace $\text{Angular}(\mathbf{g}_k, (0, 0))$ in Line 7 of Algo. 1 to make α and β have the same scale.

Consistency Loss

We propose the Consistency Loss $\mathcal{L}_{con}$ to ensure that the eyeballs in $\mathbf{x}_{fro}$ are indeed looking at $(0, 0)$ and that the head orientation is indeed the same as that of $\mathbf{x}$, further strengthening the effectiveness of the auxiliary learning of gaze frontalization. $\mathcal{L}_{con}$ includes Gaze Term $\mathcal{L}_{con}^{\mathcal{G}}$ and Head Orientation Term $\mathcal{L}_{con}^{\mathcal{H}}$, as shown in Fig. 3(d).

Gaze Term. We feed the output face image $\mathbf{x}_{fro}$ back into the Gaze Estimation Module to ensure that the output gaze direction $\hat{\mathbf{g}}_{fro}$ is $(0, 0)$:

$$\mathcal{L}_{con}^{\mathcal{G}} = \|\hat{\mathbf{g}}_{fro} - (0, 0)\|_1. \tag{7}$$

In our implementation, we replace $(0, 0)$ in Eq. 7 with the gaze label $\mathbf{g}_{ref}$ of $\mathbf{x}_{ref}$ to make the training more stable, and the loss function is changed to:

$$\mathcal{L}_{con}^{\mathcal{G}*} = \|\hat{\mathbf{g}}_{fro} - \mathbf{g}_{ref}\|_1. \tag{8}$$

We further conduct experiments to compare $\mathcal{L}_{con}^{\mathcal{G}*}$ (Eq. 8) with $\mathcal{L}_{con}^{\mathcal{G}}$ (Eq. 7) , which show that $\mathcal{L}_{con}^{\mathcal{G}*}$ is better.

Head Orientation Term. We feed $\mathbf{x}_{fro}$ back into the Gaze Estimation Module to ensure that the output head orientation $\hat{\mathbf{h}}_{fro}$ is the same as the head orientation of $\mathbf{x}$:

$$\mathcal{L}_{con}^{\mathcal{H}} = \|\hat{\mathbf{h}}_{fro} - \mathbf{h}\|_1. \tag{9}$$

Similarly to Gaze Term, in our implementation, we replace $\mathbf{h}$ in Eq. 9 with the head orientation label $\mathbf{h}_{ref}$ of $\mathbf{x}_{ref}$ to make training more stable, and the loss function is changed to:

$$\mathcal{L}_{con}^{\mathcal{H}*} = \|\hat{\mathbf{h}}_{fro} - \mathbf{h}_{ref}\|_1. \tag{10}$$

We conduct experiments to compare $\mathcal{L}_{con}^{\mathcal{H}*}$ (Eq. 10) with $\mathcal{L}_{con}^{\mathcal{H}}$ (Eq. 9), which show that $\mathcal{L}_{con}^{\mathcal{H}*}$ is better.

Total Loss Function

The total loss function of this framework is:

$$\mathcal{L} = \mathcal{L}_{\mathcal{G}} + \mathcal{L}_{\mathcal{H}} + \lambda_{fro}\mathcal{L}_{fro} + \lambda_{con}^{\mathcal{G}*}\mathcal{L}_{con}^{\mathcal{G}*} + \lambda_{con}^{\mathcal{H}*}\mathcal{L}_{con}^{\mathcal{H}*}, \tag{11}$$

Algorithm 1: Reference Image Selection

Require: Input face image $\mathbf{x}$, face images $\mathcal{D}_{ide}$ that shares the same identity as $\mathbf{x}$.
Ensure: The reference image $\mathbf{x}_{ref}$ of $\mathbf{x}$.

$s_{\min} \leftarrow +\infty$ ▷ The minimum score.
$K \leftarrow -1$ ▷ The id of the reference image in $\mathcal{D}_{ide}$.
for every candidate face image $\mathbf{x}_k \in \mathcal{D}_{ide}$ **do**
  if $\mathbf{x}_k$ equals $\mathbf{x}$ **then**
    continue ▷ $\mathbf{x}_k$ should be different from $\mathbf{x}$.
  end if
  $\alpha \leftarrow \text{Angular}(\mathbf{g}_k, (0,0))$ ▷ $\mathbf{g}_k$ is the gaze label of $\mathbf{x}_k$.
  $\beta \leftarrow \text{Angular}(\mathbf{h}_k, \mathbf{h})$ ▷ $\mathbf{h}_k$ and $\mathbf{h}$ are the head orientation labels of $\mathbf{x}_k$ and $\mathbf{x}$, respectively.
  $s \leftarrow \alpha + \beta$ ▷ The score of $\mathbf{x}_k$.
  if $s < s_{\min}$ **then** ▷ Found smaller $s_{\min}$.
    $s_{\min} \leftarrow s$
    $K \leftarrow k$
  end if
end for
$\mathbf{x}_{ref} \leftarrow \mathbf{x}_K$

where λ_{fro}, $\lambda_{con}^{\mathcal{G}*}$ and $\lambda_{con}^{\mathcal{H}*}$ are the weights used to balance different loss terms. Note that $\mathcal{L}_{\mathcal{H}}$ and $\mathcal{L}_{con}^{\mathcal{H}*}$ are optimized only when head orientation labels are provided in the training dataset. In our implementation based on image warping, we empirically set $\lambda_{fro} = 0.1$ and $\lambda_{con}^{\mathcal{G}*} = \lambda_{con}^{\mathcal{H}*} = 0.01$.

Experiments

Setup

Datasets. Following other gaze domain generalization (DG) methods (Cheng, Bao, and Lu 2022; Bao et al. 2022), we use ETH-XGaze ($\mathcal{D}_E$) (Zhang et al. 2020) and Gaze360 ($\mathcal{D}_G$) (Kellnhofer et al. 2019) for training, and MPIIGaze ($\mathcal{D}_M$) (Zhang et al. 2017b) and EyeDiap ($\mathcal{D}_D$) (Funes Mora, Monay, and Odobez 2014) for evaluation, as the former two datasets have wider gaze distributions than the latter two (Liu et al. 2021). Therefore, 4 DG tasks are formed, including $\mathcal{D}_E \rightarrow \mathcal{D}_M$, $\mathcal{D}_E \rightarrow \mathcal{D}_D$, $\mathcal{D}_G \rightarrow \mathcal{D}_M$ and $\mathcal{D}_G \rightarrow \mathcal{D}_D$. Specifically, $\mathcal{D}_E$ contains images of 80 subjects, and we use 75 of them for training and the remaining 5 for validation; hence, there is a total of 713,646 images for training, as in (Cheng, Bao, and Lu 2022). For $\mathcal{D}_G$, the training set contains 84,902 samples, where the subjects are looking at the front. For these two datasets, we use the provided pre-processed data. $\mathcal{D}_M$ contains 45,000 images for evaluation, which are normalized using (Sugano, Matsushita, and Sato 2014). For $\mathcal{D}_D$, we use the VGA videos in the screen target session and sample images every 15 frames, to obtain 16,674 images for evaluation, as in (Cheng, Bao, and Lu 2022).

Comparison Methods. We compare our proposed method with (a) Typical Gaze Estimation (TGE), (b) Gaze Domain Generalization (DG) and (c) Gaze Unsupervised Domain Adaptaion (UDA) methods. For TGE, we compare our proposed method with Full-Face (Zhang et al. 2017a), RT-Gene (Fischer, Chang, and Demiris 2018), Dilated-Net (Chen and Shi 2018), CA-Net (Cheng et al. 2020a), GazeTR (Cheng and Lu 2022) and (O Oh, Chang, and Choi 2022). They do not require target domain information when training, so it is necessary to compare them. For DG, PureGaze (Cheng, Bao, and Lu 2022) and GazeCon (Xu, Wang, and Lu 2023) have been proposed. Moreover, some gaze SDA, UDA, redirection, and unconstrained gaze estimation methods contain gaze DG modules, including FAZE (Park et al. 2019), RAT (Bao et al. 2022), CDG (Wang et al. 2022) and LatentGaze (Lee et al. 2022). We also compare our proposed gaze DG method with them as they all claimed gaze domain generalization ability. And we compare our proposed method with Baseline (Zhang et al. 2020) (based on ResNet (He et al. 2016)). Furthermore, although our proposed gaze DG method is not designed for gaze UDA, we also compare our proposed method with other gaze UDA methods for reference, including ADDA (Tzeng et al. 2017), GazeAdv (Wang et al. 2019), Gaze360 (Kellnhofer et al. 2019), DAGEN (Guo et al. 2021), PnP-GA (Liu et al. 2021), RUDA (Bao et al. 2022), CRGA (Wang et al. 2022) and LatentGaze (Lee et al. 2022). Note that gaze UDA methods need some unlabeled target domain samples for adaptation, while gaze DG methods do not need these target domain samples for training.

Gaze Redirection Methods. Although GFAL framework is designed for the gaze DG task and not the gaze redirection task, it has some similarities to gaze redirection. Therefore, we conduct experiments based on representative gaze redirection methods, including DeepWarp (Ganin et al. 2016), FAZE (Park et al. 2019), ST-ED (Zheng et al. 2020) and CUDA-GHR (Jindal and Wang 2023).

Implementation Details. We use NVIDIA GPUs for the experiments. We use ResNet (He et al. 2016) as the backbone. Models are trained for 25 epochs for $\mathcal{D}_E$ and 100 epochs for $\mathcal{D}_G$, using a batch size of 50. The Adam optimizer is used with $lr = 10^{-4}$, $\beta_1 = 0.9$ and $\beta_2 = 0.95$. All input images are of the size 224×224 and are normalized to $[0, 1]$.

Furthermore, since CDG (Wang et al. 2022) is designed based on data augmentation, including a random color field and grayscale, we apply these data augmentation strategies to GFAL framework and compare GFAL framework with CDG. The results in the 5^{th} row show that our framework can also surpass CDG (Wang et al. 2022).

Comparison with SOTA Methods

To show the superior cross-domain gaze estimation performance of GFAL framework, we compare our proposed method with SOTA methods in the DG and UDA tasks.

Domain Generalization. The results are shown in the 3^{nd}-6^{th} rows in Table 1. For a fair comparison, we categorize the gaze DG methods according to different backbones and whether data augmentation is performed.

Type	Methods	$\lvert\mathcal{D}_t\rvert$	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
TGE	Full-Face	0	12.35	30.15	11.13	14.42
	RT-Gene	0	-	-	21.81	38.60
	Dilated-Net	0	-	-	18.45	23.88
	CA-Net	0	-	-	27.13	31.41
	Oh *et al.*	0	-	-	8.31	7.10
	GazeTR	0	8.70	10.98	7.96	8.88
	Ours (Res-18)†	0	**5.72**	6.97	**7.18**	7.38
	Ours (Res-50)‡	0	6.09	**6.48**	7.23	**6.76**
DG†	Baseline†	0	7.48	7.86	7.70	8.12
	FAZE†	0	9.49	21.03	-	-
	PureGaze†	0	9.14	8.37	9.28	9.32
	RAT†	0	7.92	8.65	7.60	8.16
	LatentGaze†	0	7.98	9.81	-	-
	GazeCon†	0	6.89	7.78	7.82	8.52
	Ours (Res-18)†	0	**5.72**	**6.97**	**7.18**	**7.38**
DG‡	Baseline‡	0	7.11	7.38	7.38	7.17
	FAZE‡	0	7.80	15.85	-	-
	PureGaze‡	0	7.08	7.48	7.62	7.70
	RAT‡	0	7.40	8.38	7.69	8.32
	GazeCon‡	0	8.35	8.80	8.24	8.83
	Ours (Res-50)‡	0	**6.09**	**6.48**	**7.23**	**6.76**
DG$^{\dagger *}$	GazeCon$^{\dagger *}$	0	6.50	7.44	7.55	9.03
	Ours (Res-18)$^{\dagger *}$	0	**6.47**	**7.04**	**6.98**	**9.02**
DG$^{\ddagger *}$	CDG$^{\ddagger *}$	0	6.73	7.95	7.03	7.27
	GazeCon$^{\ddagger *}$	0	7.20	8.76	8.22	11.15
	Ours (Res-50)$^{\ddagger *}$	0	**5.68**	**6.23**	**6.29**	**6.69**
UDA† (for reference)	ADDA†	500	6.65	8.24	6.27	9.53
	GazeAdv†	100	6.36	7.62	7.54	8.43
	Gaze360†	100	6.24	7.47	7.17	7.66
	DAGEN†	500	5.73	6.77	7.38	8.00
	PnP-GA†	10	5.53	5.87	6.18	7.92
	RUDA†	100	5.70	7.52	6.20	7.02
	LatentGaze†	<100	5.21	7.81	-	-
	Ours (Res-18)† **(w/o $\mathcal{D}_t$)**	**0**	5.72 **(4/8)**	6.97 **(3/8)**	7.18 **(5/7)**	7.38 **(2/7)**

Table 1: Results of SOTA gaze estimation methods. Angular error in degrees are shown. $\lvert\mathcal{D}_t\rvert$ indicates target domain sample numbers for training. † and ‡ indicate that the backbone is ResNet-18 and 50, respectively. * indicates that data augmentation is used.

Typical gaze estimation methods are usually trained and tested on the same domain and do not consider the cross-dataset setting. As a result, their performances on $\mathcal{D}_t$ are poor. In contrast, GFAL framework surpasses all the typical gaze estimation methods (2nd row in Tab. 1).

In addition, we compare GFAL framework with other SOTA gaze DG methods. As shown in the 3rd and 4th rows in Tab. 1, GFAL framework can outperform the other methods by a large margin. Note that ResNet-50 may not always yield better results than ResNet-18. This is because deeper networks like ResNet-50 may be more prone to overfit the training data (Schmidt 2023).

Furthermore, since CDG and GazeCon is designed based on data augmentation, including a random color field and grayscale, we apply these data augmentation strategies to GFAL framework and compare GFAL framework with them. The results in the 5th and 6th row show that our framework can also surpass them.

Moreover, our proposed GFAL shows good scalability across datasets. GFAL can achieve good domain generalization performance while using different training datasets, where the largest contains 713,646 images, while the smallest contains 84,902 images.

Unsupervised Domain Adaptation. We compare GFAL framework with SOTA gaze UDA methods for reference. The results are shown in the last two rows in Tab. 1. Our proposed method (ResNet-18) can surpass 4 of 7, 5 of 7, 2 of 6, and 5 of 6 methods in the $\mathcal{D}_E \rightarrow \mathcal{D}_M$, $\mathcal{D}_E \rightarrow \mathcal{D}_D$, $\mathcal{D}_G \rightarrow \mathcal{D}_M$ and $\mathcal{D}_G \rightarrow \mathcal{D}_D$ tasks, respectively.

Task	FAZE	ST-ED	CUDA-GHR	Baseline†	Ours†
$\mathcal{D}_E \rightarrow \mathcal{D}_M$	8.17	7.30	7.58	7.48	**5.72**
$\mathcal{D}_E \rightarrow \mathcal{D}_D$	11.61	8.14	8.99	7.86	**6.97**

Table 2: Cross-dataset validation results of different gaze redirection methods on the gaze estimation task. The results show the angular error in degrees.

Relation to Gaze Redirection Methods

Comparison with Gaze Redirection Methods. This experiment aims to prove that GFAL can surpass gaze redirection methods in the cross-domain gaze estimation task.

We first pretrain gaze redirection models using $\mathcal{D}_E$ (Zhang et al. 2020). Then we use these models to generate $\lvert\mathcal{D}_s\rvert$ face images with random gaze directions, as in (Zheng et al. 2020), forming the synthetic training set $\mathcal{D}_{syn}$, where $\lvert\mathcal{D}_s\rvert$ is the size of the source domain training set. Then, we train the gaze estimation model using the ResNet-18 (He et al. 2016) backbone and the training set $\mathcal{D}_s \cup \mathcal{D}_{syn}$. Note that we do not use $\mathcal{D}_G$ (Kellnhofer et al. 2019), as $\mathcal{D}_G$ does not provide head orientation labels, and these gaze redirection methods need head orientation labels for training. The results are shown in Tab. 2.

It can be concluded that GFAL framework is superior to SOTA gaze redirection methods. The reason behind this is that these models still suffer from the infinite and continuous labels problem, although the training dataset is extended and enriched so that it contains more gaze information. Conversely, GFAL framework can achieve better performance than gaze redirection methods.

Plug Gaze Redirection Methods into GFAL Framework. This experiment aims to prove that gaze redirection methods can be plugged into GFAL framework to improve cross-domain gaze estimation performance.

Here, we plug the SOTA gaze redirection methods into GFAL framework to replace the image warping procedure and use the gaze feature $\mathbf{z}$ only to generate redirected face images that look at $(0, 0)$. All the experiments are conducted with a ResNet-18 (He et al. 2016) backbone for extracting $\mathbf{z}$.

The results are shown in Tab. 3. It can be seen that plugging these gaze redirection methods into GFAL framework can improve gaze domain generalization performance as

Strategy	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$	Avg
Baseline†	7.48	7.86	7.70	8.12	7.79
Ours (DeepWarp)†	7.35	8.93	**6.73**	**6.89**	7.48
Ours (FAZE)†	8.10	9.37	6.87	7.66	8.00
Ours (ST-ED)†	7.22	8.20	7.40	8.15	7.74
Ours (Warping)†	**5.72**	**6.97**	7.18	7.38	**6.81**

Table 3: Cross-dataset validation results of plugging different SOTA gaze redirection methods into our proposed framework. The results show the angular error in degrees.

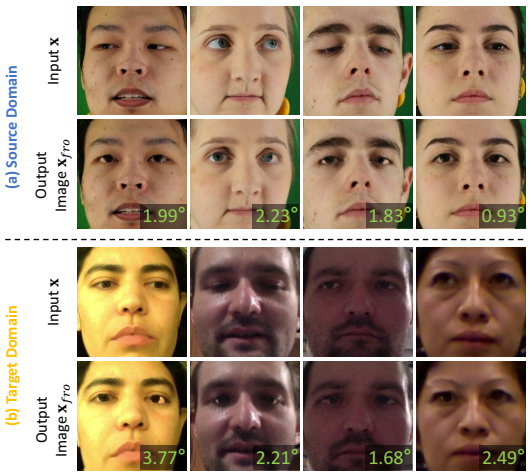

Figure 4: Visualization of $\mathbf{x}_{fro}$. Angular errors between $\hat{\mathbf{g}}_{fro}$ and $(0, 0)$ are shown in the lower right corner of $\mathbf{x}_{fro}$.

well as the image warping for most of time. However, the overall (average) performance of the implementation based on gaze redirection methods cannot surpass that the implementation based on image warping. This is because SOTA gaze redirection networks are complex and difficult to train, leading to difficulty in training gaze estimation models based on these kinds of implementation.

Gaze Frontalization Visualization

This experiment aims to show that the extracted feature $\mathbf{z}$ can correctly perform the gaze frontalization process, *i.e.*, rotating the eyeball in $\mathbf{x}$ to $(0, 0)$. We visualize $\mathbf{x}_{fro}$ according to $\mathbf{x}$, results are shown in Fig. 4. The eyeball directions in $\mathbf{x}$ are successfully rotated from (θ, ψ) to $(0, 0)$ in the output image $\mathbf{x}_{fro}$. The angular error between $\hat{\mathbf{g}}_{fro} = \mathcal{G}(\mathcal{F}(\mathbf{x}_{fro}))$ and $(0, 0)$ shown in the lower right corner of every $\mathbf{x}_{fro}$ in Fig. 4 is small enough to support this observation. This observation indicates that our proposed framework can successfully force $\mathbf{z}$ to process gaze frontalization.

Ablation Study

Comparison of Different Choices of $\mathcal{L}_{fro}$. We conduct an experiment to compare different choices of $\mathcal{L}_{fro}$, including

Strategy	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$	Avg
L_1	7.28	7.40	6.76	7.47	7.23
L_2	7.12	8.25	7.18	8.19	7.69
SSIM	6.78	8.59	**6.33**	7.56	7.32
MS-SSIM	**5.72**	**6.97**	7.18	**7.38**	**6.81**
$\mathcal{L}_{con}$	6.02	8.16	7.91	7.47	7.39
$\mathcal{L}^*_{con}$	**5.72**	**6.97**	**7.18**	**7.38**	**6.81**
$(0, 0) \rightarrow (\theta, \psi)$	7.33	7.15	7.56	7.87	7.48
$(\theta, \psi) \rightarrow (0, 0)$	**5.72**	**6.97**	**7.18**	**7.38**	**6.81**
Baseline	7.48	7.86	7.70	8.12	7.79
$\mathcal{L}_{fro}$	7.92	8.92	7.41	8.59	8.21
$\mathcal{L}_{fro} + \mathcal{L}_{con}$	**5.72**	**6.97**	**7.18**	**7.38**	**6.81**

Table 4: Ablation study results of different strategies. The results show the angular error in degrees.

distance metrics L_1, L_2, SSIM (Wang et al. 2004) and MS-SSIM (Wang, Simoncelli, and Bovik 2003). The results are shown in the second row of Tab. 4, which indicate that using MS-SSIM can achieve the best performance and is the best choice for the distance metric between $\mathbf{x}_{ref}$ and $\mathbf{x}_{fro}$.

Comparison of $\mathcal{L}_{con}$ and $\mathcal{L}^*_{con}$. We use $\mathcal{L}^*_{con}$ (Eq. 8 and 10) instead of $\mathcal{L}_{con}$ (Eq. 7 and 9) for more stable training. Here, we compare these two types of loss terms, and the results are shown in the third row of Tab. 4. It is indicated that using $\mathcal{L}^*_{con}$ is better for Consistency Loss than using $\mathcal{L}_{con}$.

Comparison of $(0, 0) \rightarrow (\theta, \psi)$ and $(\theta, \psi) \rightarrow (0, 0)$. We use the $(\theta, \psi) \rightarrow (0, 0)$ strategy (rotating (θ, ψ) to $(0, 0)$) instead of $(0, 0) \rightarrow (\theta, \psi)$ to make every $\mathbf{g}$ have a clear and unique target $(0, 0)$, further constraining the gaze estimation learning. The comparison results of these 2 strategies are shown in the fourth row of Tab. 4, and the results show that $(\theta, \psi) \rightarrow (0, 0)$ is better.

Contributions of the Loss Terms. Results are shown in the last row of Tab. 4. Using only $\mathcal{L}_{fro}$ cannot improve the cross-domain gaze estimation performance. Furthermore, using both $\mathcal{L}_{fro}$ and $\mathcal{L}_{con}$ can significantly improve performance. This further shows the effectiveness of both Gaze Frontalization Module and Consistency Loss for gaze frontalization auxiliary learning.

Conclusion

In this paper, we propose a novel gaze domain generalization (DG) framework GFAL, which aims to utilize the embedding of gaze frontalization process to improve cross-domain gaze estimation performance without any target domain information during training. Experimental results show that GFAL framework can achieve SOTA performance on gaze DG task, which is competitive with or even superior to the SOTA gaze UDA methods, and surpass most of representative gaze redirection methods. Moreover, various types of implementations can be plugged into GFAL framework, which can improve cross-domain gaze estimation performance for most of time. This work provides new insights for cross-domain gaze estimation. In the future, we will extend the gaze frontalization methods, such as 3D reconstruction.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China (NSFC) under Grant 62372019, and partially supported by Peng Cheng Laboratory (PCL2023A10-2).

References

Bao, Y.; Cheng, Y.; Liu, Y.; and Lu, F. 2021. Adaptive feature fusion network for gaze tracking in mobile tablets. In *2020 25th international conference on pattern recognition*, 9936–9943. 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.

Chen, Z.; and Shi, B. 2020. Offset calibration for appearance-based gaze estimation via gaze decomposition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 270–279.

Chen, Z.; and Shi, B. E. 2018. Appearance-based gaze estimation using dilated-convolutions. In *Proceedings of the asian conference on computer vision*, 309–324. Springer.

Chen, Z.; and Shi, B. E. 2022. Towards high performance low complexity calibration in appearance based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 1174–1188.

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. 2020a. 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*, 3341–3347. IEEE.

Cheng, Y.; Lu, F.; and Zhang, X. 2018. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the european conference on computer vision*, 100–115.

Cheng, Y.; Zhang, X.; Lu, F.; and Sato, Y. 2020b. Gaze estimation by exploring two-eye asymmetry. *IEEE transactions on image processing*, 29: 5259–5272.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International conference on learning representations*.

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*, 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*, 255–258.

Ganin, Y.; Kononenko, D.; Sungatullina, D.; and Lempitsky, V. 2016. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *Proceedings of the european conference on computer vision*, 311–326. Springer.

Guo, Z.; Yuan, Z.; Zhang, C.; Chi, W.; Ling, Y.; and Zhang, S. 2021. Domain Adaptation Gaze Estimation by Embedding with Prediction Consistency. In *Proceedings of the asian conference on computer vision*, 292–307. Springer.

He, J.; Pham, K.; Valliappan, N.; Xu, P.; Roberts, C.; Lagun, D.; and Navalpakkam, V. 2019. On-device few-shot personalization for real-time gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*.

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.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; and Kavukcuoglu, K. 2015. Spatial transformer networks. In *Proceedings of the 28th international conference on neural information processing systems*, 2017–2025.

Jindal, S.; and Wang, X. E. 2023. Cuda-ghr: Controllable unsupervised domain adaptation for gaze and head redirection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 467–477.

Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; and Torralba, A. 2019. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6912–6921.

Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; and Torralba, A. 2016. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2176–2184.

Lahiri, A.; Agarwalla, A.; and Biswas, P. K. 2018. Unsupervised domain adaptation for learning eye gaze from a million synthetic images: An adversarial approach. In *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing*, 1–9.

Lee, I.; Yun, J.-S.; Kim, H. H.; Na, Y.; and Yoo, S. B. 2022. LatentGaze: Cross-Domain Gaze Estimation through Gaze-Aware Analytic Latent Code Manipulation. In *Proceedings of the asian conference on computer vision*, 3379–3395.

Liu, G.; Yu, Y.; Mora, K. A. F.; and Odobez, J.-M. 2018. A differential approach for gaze estimation with calibration. In *British machine vision conference*, volume 2, 6.

Liu, G.; Yu, Y.; Mora, K. A. F.; and Odobez, J.-M. 2019. A differential approach for gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 43(3): 1092–1099.

Liu, R.; Bao, Y.; Xu, M.; Wang, H.; Liu, Y.; and Lu, F. 2022. Jitter Does Matter: Adapting Gaze Estimation to New Domains. arXiv:2210.02082.

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.

Lu, F.; Sugano, Y.; Okabe, T.; and Sato, Y. 2011. Inferring human gaze from appearance via adaptive linear regression. In *Proceedings of the IEEE international conference on computer vision*, 153–160.

Lu, F.; Sugano, Y.; Okabe, T.; and Sato, Y. 2014. Adaptive Linear Regression for Appearance-Based Gaze Estimation. *IEEE transactions on pattern analysis and machine intelligence*, 36(10): 2033–2046.

MacQueen, J. 1967. Classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability*, 281–297.

Mole, C.; Pekkanen, J.; Sheppard, W. E.; Markkula, G.; and Wilkie, R. M. 2021. Drivers use active gaze to monitor waypoints during automated driving. *Scientific Reports*, 11(1): 1–18.

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.

Park, S.; Mello, S. D.; Molchanov, P.; Iqbal, U.; Hilliges, O.; and Kautz, J. 2019. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9368–9377.

Park, S.; Spurr, A.; and Hilliges, O. 2018. Deep pictorial gaze estimation. In *Proceedings of the european conference on computer vision*, 721–738.

Park, S.; Zhang, X.; Bulling, A.; and Hilliges, O. 2018. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM symposium on eye tracking research and applications*, 1–10.

Schmidt, J. 2023. Testing for Overfitting. arXiv:2305.05792.

Sugano, Y.; Matsushita, Y.; and Sato, Y. 2014. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1821–1828.

Sun, L.; Liu, Z.; and Sun, M.-T. 2015. Real time gaze estimation with a consumer depth camera. *Information Sciences*, 320: 346–360.

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.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5265–5274.

Wang, K.; Zhao, R.; Su, H.; and Ji, Q. 2019. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11907–11916.

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.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multiscale structural similarity for image quality assessment. In *Thirty-seventh asilomar conference on signals, systems and computers 2003*, volume 2, 1398–1402. IEEE.

Wang, Z.; Zhao, Y.; and Lu, F. 2022. Gaze-Vergence-Controlled See-Through Vision in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11): 3843–3853.

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.

Yu, Y.; Liu, G.; and Odobez, J.-M. 2019a. Deep Multitask Gaze Estimation with a Constrained Landmark-Gaze Model. In *Proceedings of the european conference on computer vision workshops*, 456–474. Springer.

Yu, Y.; Liu, G.; and Odobez, J.-M. 2019b. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11937–11946.

Yu, Y.; and Odobez, J.-M. 2020. Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7314–7324.

Zhang, X.; Park, S.; Beeler, T.; Bradley, D.; Tang, S.; and Hilliges, O. 2020. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Proceedings of the european conference on computer vision*, 365–381. Springer.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017a. 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. 2017b. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1): 162–175.

Zheng, Y.; Park, S.; Zhang, X.; De Mello, S.; and Hilliges, O. 2020. Self-learning transformations for improving gaze and head redirection. *Advances in neural information processing systems*, 33: 13127–13138.