Learning a Generalized Gaze Estimator from Gaze-Consistent Feature

Mingjie Xu[1], **Haofei Wang**[2], **Feng Lu**[1, 2*]

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

Abstract

Gaze estimator computes the gaze direction based on face images. Most existing gaze estimation methods perform well under within-dataset settings, but can not generalize to unseen domains. In particular, the ground-truth labels in unseen domain are often unavailable. In this paper, we propose a new domain generalization method based on gaze-consistent features. Our idea is to consider the gaze-irrelevant factors as unfavorable interference and disturb the training data against them, so that the model cannot fit to these gaze-irrelevant factors, instead, only fits to the gaze-consistent features. To this end, we first disturb the training data via adversarial attack or data augmentation based on the gaze-irrelevant factors, *i.e*., identity, expression, illumination and tone. Then we extract the gaze-consistent features by aligning the gaze features from disturbed data with non-disturbed gaze features. Experimental results show that our proposed method achieves state-of-the-art performance on gaze domain generalization task. Furthermore, our proposed method also improves domain adaption performance on gaze estimation. Our work provides new insight on gaze domain generalization task.

Introduction

Gaze indicates where a person is looking. It has been used in a variety of applications, such as human-computer interaction (Katsini et al. 2020; Wang et al. 2015), virtual reality (Wang, Zhao, and Lu 2022; Zhang, Wang, and Shi 2021), and saliency prediction (Xu, Sugano, and Bulling 2016). Gaze estimation techniques compute the gaze direction, which can be roughly categorized into the model-based approaches and the appearance-based approaches. Model-based approaches usually require dedicated devices, *i.e*., eye trackers, while appearance-based approaches only use inexpensive webcams and their performance is also relatively poor (Lu et al. 2011, 2014; Zhang et al. 2017a).

With the advances of machine learning techniques, appearance-based gaze estimation has achieved significant progress (Zhang et al. 2017a; Cheng et al. 2020a). Appearance-based gaze estimation approaches usually use hand-crafted gaze features (Lu et al. 2011, 2014) or features extracted using deep learning models (Zhang et al. 2017a; Cheng et al. 2020a) to regress gaze directions. They usually trained and tested on the same dataset, *i.e*., the same domain. However, these algorithms suffer from severe performance degradation when they are trained on one dataset (source domain) and tested on another dataset (target domain). Such problem has been defined as cross-domain gaze estimation problem (Liu et al. 2021). This limits its applications, since the background is often changing in real-world settings.

Cross-domain gaze estimation is more challenging, as it relies on gaze feature extracted from tiny eyeball regions on the face images. In fact, a number of domain adaption methods have been proposed for other computer vision tasks, *e.g*., object recognition (Ghifary, Kleijn, and Zhang 2014). For cross-domain gaze estimation, Liu *et al*. proposed the first plug-and-play gaze adaptation framework (PnP-GA), which is an ensemble of networks that learn collaboratively with the guidance of outliers (Liu et al. 2021).

At the core of the cross-domain gaze estimation is to balance the model generalization ability and over-fitting problem. On the one hand, more diversified samples are favored so that the trained model is more generalized to different conditions; on the other hand, more data on the same domain may lead to over-fitting. Therefore, our goal is to disturb the data with diverse gaze-irrelevant factors while keeping the gaze feature consistent. We define several gaze-irrelevant factors: identity, facial expression, illumination and tone.

In this paper, our motivation is to consider the gaze-irrelevant features as unfavorable interference and disturb the training data against them. We argue that by introducing more diversified training data disturbed by these factors, the gaze model fails to fit to the gaze-irrelevant factors, rather, it only fits to the gaze-consistent features. Towards this end, we first synthesize diversified training data based on gaze-irrelevant factors using adversarial attack or data augmentation. Then we learn the gaze-consistent feature by minimizing the distance between the gaze features from disturbed data and non-disturbed gaze features. The primary contributions of this paper are:

- We propose a generalized gaze estimator that learns the gaze direction from gaze-consistent features for gaze domain generalization task. We first disturb the training data based on gaze-irrelevant factors (identity, facial expression, illumination and tone), and then extract the gaze-consistent features to learn the gaze direction.

*Corresponding Author.

- Experimental results show that our proposed method achieves state-of-the-art performance on gaze domain generalization task, and is able to indeed confuse gaze-irrelevant factors. Moreover, our proposed method can also improve gaze domain adaption performance.

Related Work

Typical Gaze Estimation Typical appearance-based gaze estimation uses Convolutional Neural Network (CNN) to regress gaze direction based on face images (Cheng et al. 2021) or eye images (Zhang et al. 2015), which has been first proposed in (Zhang et al. 2015, 2017b). Following this work, a number of gaze estimation methods were proposed (Zhang et al. 2017a; Cheng, Lu, and Zhang 2018; Cheng et al. 2020b; Bao et al. 2021; Park, Spurr, and Hilliges 2018; Cheng et al. 2020a; Fischer, Chang, and Demiris 2018). For example, Park *et al*. developed a pictorial representation for modelling the pupil center and eyeball center (Park, Spurr, and Hilliges 2018). Cheng *et al*. (Cheng, Lu, and Zhang 2018; Cheng et al. 2020b) used a pair of eye images and exploited the asymmetry property to estimate gaze. However, the above-mentioned methods only estimate gaze in a within-dataset setting, they often experience a significant performance degradation when tested on unseen domain, *i.e*., new dataset. The cross-domain gaze estimation task is still challenging.

Unsupervised Domain Adaption Unsupervised Domain Adaption (UDA) adapts the model to target domain using a few unlabeled target domain samples. Kellnhofer *et al*. (Kellnhofer et al. 2019) used domain discriminator and left-right symmetry to adapt gaze estimation model to target domain. Wang *et al*. (Wang et al. 2019) proposed a UDA method on gaze estimation using adversarial learning and Bayesian inference. Guo *et al*. (Guo et al. 2020) introduced a new target domain representation embedding mechanism with prediction consistency for gaze UDA. Liu *et al*. (Liu et al. 2021) used network ensemble with momentum and outlier guidance to achieve gaze UDA. Besides, Bao *et al*. (Bao et al. 2022) proposed a new gaze UDA approach based on rotation consistency. Wang *et al*. (Wang et al. 2022) proposed contrastive regression for gaze UDA. Nevertheless, all these methods require target domain information, which is usually difficult to obtain in real-world settings.

Domain Generalization Unlike UDA, Domain Generalization (DG) forbids access to any target domain information when training the model on source domain, which is more difficult. Domain generalization tasks are often tackled by feature distance minimization (Li et al. 2018), feature disentanglement (Ilse et al. 2020), data generation (Tobin et al. 2017), etc. To the best of our knowledge, PureGaze (Cheng, Bao, and Lu 2022) is the first and the only domain generalization method developed for gaze estimation. They proposed a gaze feature purification mechanism based on adversarial learning to eliminate gaze-irrelevant factors. Although PureGaze improved cross-dataset performance, they did not analyze and explain the reason why these gaze-irrelevant factors lead to performance degradation.

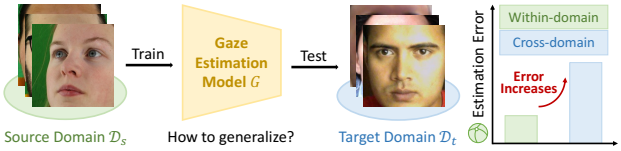


Figure 1: Illustration of the gaze domain generalization task.

Domain Generalization for Gaze Estimation

Preliminaries

We first give preliminaries of the gaze estimation and domain generalization task. The source domain dataset is defined as $\mathcal{D}_s = \left\{(\mathbf{x}_i^s, \mathbf{y}_i^s)|_{i=1}^{N_s}\right\}$, where $\mathbf{x}_i^s$ denotes the i-th input image in source domain, $\mathbf{y}_i^s$ is the corresponding gaze label $(pitch, yaw)$, and N_s is the total number of images. Similarly, the target domain dataset is defined as $\mathcal{D}_t = \left\{(\mathbf{x}_i^t, \mathbf{y}_i^t)|_{i=1}^{N_t}\right\}$. The gaze estimation method $G(\cdot) = R(F(\cdot))$ is usually trained **on the source domain** $\mathcal{D}_s$, where F represents the feature extractor and R denotes the Multi Layer Perceptron (MLP) that regresses feature vectors to gaze directions. Given an input face image $\mathbf{x}$, the model outputs the predicted gaze direction $\hat{\mathbf{y}} = G(\mathbf{x})$.

As shown in Fig. 1, the goal of cross-domain gaze estimation task is to make the predictions with the data from the target domain $\mathcal{D}_t$ using the model G trained on the source domain $\mathcal{D}_s$. The prediction error is defined as the angular error between the predicted gaze direction $G(\mathbf{x})$ and the ground truth gaze label $\mathbf{y}$:

$$E[G(\mathbf{x}), \mathbf{y}] = \frac{G(\mathbf{x}) \cdot \mathbf{y}}{\|G(\mathbf{x})\| \cdot \|\mathbf{y}\|}. \tag{1}$$

If we know the target domain data $\{(\mathbf{x}_i^t, \mathbf{y}_i^t)\}$, we can easily fine-tune the model G by updating its parameters θ:

$$\theta = \arg\min_{\theta} \sum\nolimits_i E[G(\mathbf{x}_i^t), \mathbf{y}_i^t]. \tag{2}$$

However, in practice we may not know the target domain in advance. In such cases, we need to optimize θ without using the above minimization and target domain data $\{(\mathbf{x}_i^t, \mathbf{y}_i^t)\}$. This defines the problem of domain generalization.

Why Is It Hard?

Based on the above definition, one of the biggest challenge is that the CNN model is prone to overfitting on the source domain without known the target domain. As a result, the model generalizability and performance drop, and the trained model performs poor on the unseen target domain data, as shown in the right part of Fig. 2. To tackle such a problem, a common approach in computer vision is to improve the diversity of source domain training dataset.

However, the uniqueness of gaze estimation task makes it a more challenging task. As shown in Fig. 2, the orientation of eyeball determines the gaze direction, unfortunately, the eyeball is only a tiny region on the facial image while other regions contains a plethora of disturbing factors, *e.g*., various facial expressions and illuminations. Thus, even though

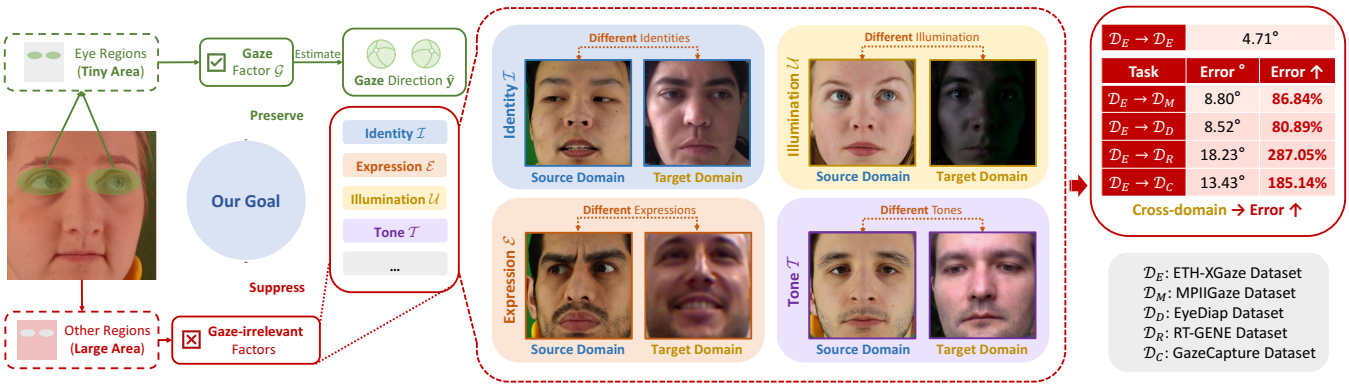


Figure 2: Illustration of our observation and motivation. We observed the uniqueness of gaze estimation task and domain gap in gaze estimation. The model generalizability reduces as the extracted features are easily overfitted to gaze-irrelevant factors.

the diversity of gaze is sufficient to train the deep neural network (*e.g.*, the ETH-XGaze (Zhang et al. 2020) dataset contains about 757 thousand samples for learning 2 degree-of-freedom gaze direction), the extracted features are easily overfitted to gaze-irrelevant factors, which reduces the model generalizability.

Formally, we denote the gaze feature extracted on source domain as $\mathbf{z}^s$, *i.e.*, $\mathbf{z}^s = F(\mathbf{x}^s)$. Based on the above analysis, $\mathbf{z}^s$ contains gaze factor ($\mathcal{G}^s$) and gaze-irrelevant factors (*e.g.*, **identity** $\mathcal{I}^s$, **expression** $\mathcal{E}^s$, **illumination** $\mathcal{U}^s$, and **tone** $\mathcal{T}^s$ as shown in Fig. 2), which can be denoted as $\mathbf{z}^s \leftarrow \{\mathcal{G}^s, \mathcal{I}^s, \mathcal{E}^s, \mathcal{U}^s, \mathcal{T}^s, ...\}$, where $\leftarrow$ indicates that these factors are combined in a way that feature computation need. Similarly, the gaze feature extracted on target domain can be expressed as $\mathbf{z}^t \leftarrow \{\mathcal{G}^t, \mathcal{I}^t, \mathcal{E}^t, \mathcal{U}^t, \mathcal{T}^t, ...\}$. Note that we do not assume that different gaze-irrelevant factors are independent.

For $\mathbf{x}^s$ from the source domain and $\mathbf{x}^t$ from target domain which have similar gaze, the gaze features $\mathbf{z}^s = F(\mathbf{x}^s)$ and $\mathbf{z}^t = F(\mathbf{x}^t)$ are similar in gaze factor $\mathcal{G}$ but dissimilar in gaze-irrelevant factors like $\mathcal{I}$, $\mathcal{E}$, $\mathcal{U}$, $\mathcal{T}$, etc. When the gaze estimation model G is trained on source domain $\mathcal{D}_s$ only, the extracted feature $\mathbf{z}^s$ will be easily overfitted to gaze-irrelevant factors, such as $\mathcal{I}^s, \mathcal{E}^s, \mathcal{U}^s, \mathcal{T}^s$. When tested G on target domain $\mathcal{D}_t$, we will obtain gaze-irrelevant factors $\mathcal{I}^t, \mathcal{E}^t, \mathcal{U}^t, \mathcal{T}^t$ that are different to $\mathcal{I}^s, \mathcal{E}^s, \mathcal{U}^s, \mathcal{T}^s$. As a result, the predicted gaze directions $\hat{\mathbf{y}}^t$ becomes less accurate, which leads to significant performance drop and error increase in target domain, as shown in the right part of Fig. 2.

Therefore, the following two prerequisites are necessary to address the above problem:

- **Prerequisite 1**: try to disturb the information from gaze-irrelevant factors ($\mathcal{I}, \mathcal{E}, \mathcal{U}, \mathcal{T}$) in the dataset so that they can not be fitted, no matter these gaze-irrelevant factors are coupled or disentangled.
- **Prerequisite 2**: make sure that the gaze information $\mathcal{G}$ in the dataset is correct and can be learned, *i.e.*, disturbing gaze-irrelevant features does not affect gaze-relevant feature in terms of performing gaze estimation task.

Note that these gaze-irrelevant factors are chosen simply because they are the most obvious factors to distinguish between different gaze datasets.

Our Idea: Gaze-Irrelevant Feature Disturbation

Following the above two prerequisites, we propose our technology roadmap.

For Prerequisite 1, we disturb the training data from specific gaze-irrelevant factors (identity $\mathcal{I}$, expression $\mathcal{E}$, illumination $\mathcal{U}$ and tone $\mathcal{T}$) by employing the adversarial attack or data augmentation approach to synthesize a pair/group of data. These synthesized data are conflicting so that they can not fit to gaze-irrelevant factors. We name this phase as the Gaze-irrelevant Feature Disturbation Phase (**GiFD Phase**, see Fig. 3(I)).

For Prerequisite 2, the above synthesized data are fed into the gaze model and make sure that the extracted gaze features are consistent on gaze factor $\mathcal{G}$. We call this phase Gaze-consistent Feature Extraction Phase (**GcFE Phase**, see Fig. 3(II)).

In this way, we guarantee that the extracted features can fit to gaze factor $\mathcal{G}$ and cannot fit to the gaze-irrelevant factors, such as $\mathcal{I}, \mathcal{E}, \mathcal{U}$ and $\mathcal{T}$.

Learning from Gaze-Consistent Feature

Gaze & Gaze-Irrelevant Tasks

We use ResNet-18 (He et al. 2016) as the backbone to construct multiple branches to process gaze factor $\mathcal{G}$ and other gaze-irrelevant factors ($\mathcal{I}$, $\mathcal{E}$ or $\mathcal{U}$). These branches are used as shown in Fig. 3. All the branches share weights on `layer1`, and the subsequent layers are task specific. At inference time, only gaze branch is used for gaze estimation. Details are described below.

- **Gaze** ($\mathcal{G}$): The gaze branch regresses the gaze direction. We use L1 Loss as gaze loss function:

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

  where $\mathbf{y}$ is the ground truth gaze direction and $\hat{\mathbf{y}}$ is the estimated gaze direction.
- **Identity** ($\mathcal{I}$): The identity branch classifies the identity label using the commonly-used face recognition algorithm CosFace (Wang et al. 2018). Note that ETH-XGaze (Zhang et al. 2020) provides identity labels. For Gaze360 (Kellnhofer et al. 2019), we first use

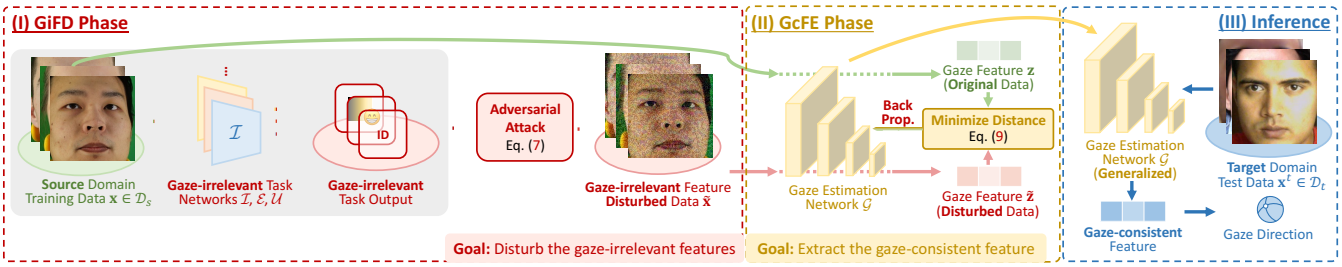


Figure 3: Illustration of the roadmap. (I) GiFD Phase, (II) GcFE Phase, (III) Inference Procedure.

CosFace (Wang et al. 2018) pretrained on CASIA-WebFace (Yi et al. 2014) to generate face feature vectors, then use k-means clustering (MacQueen 1967) to obtain pseudo identity labels. We use Focal Loss (Lin et al. 2017) as identity loss function:

$$\mathcal{L}_{ide} = \mathrm{FL}(\mathbf{y}_{ide}, \hat{\mathbf{y}}_{ide}), \tag{4}$$

where FL denotes the Focal Loss, $\mathbf{y}_{ide}$ denotes the ground truth (pseudo label for Gaze360) identity label and $\hat{\mathbf{y}}_{ide}$ denotes the estimated identity label.

- **Expression** ($\mathcal{E}$): The expression branch classifies the expression label. It consists of a feature extractor and a MLP (based on CosFace (Wang et al. 2018)), which is commonly-used in classification tasks (He et al. 2016). We use 7-class expression classifier pretrained on the FER2013 dataset (Goodfellow et al. 2013) and VGG-19 (Simonyan and Zisserman 2015) backbone to obtain pseudo expression labels. We use Focal Loss (Lin et al. 2017) as expression loss function:

$$\mathcal{L}_{exp} = \mathrm{FL}(\mathbf{y}_{exp}, \hat{\mathbf{y}}_{exp}), \tag{5}$$

where $\mathbf{y}_{exp}$ is the pseudo expression label and $\hat{\mathbf{y}}_{exp}$ is the estimated expression label.

- **Illumination** ($\mathcal{U}$): The illumination branch regresses the average illumination intensity of one image, which consists of a feature extractor and a MLP. We use the Value channel in the HSV (Hue, Saturation and Value) color space to approximate the illumination intensity (Hu, Song, and Zhi 2010). We use L1 Loss as illumination loss function:

$$\mathcal{L}_{illu} = \|\mathbf{y}_{illu} - \hat{\mathbf{y}}_{illu}\|_1, \tag{6}$$

where $\mathbf{y}_{illu}$ is the ground truth illumination intensity (Value channel in HSV color space) and $\hat{\mathbf{y}}_{illu}$ is estimated illumination intensity.

- **Tone** ($\mathcal{T}$): Tone indicates the level of image saturation. Note that here we do not use one branch to estimate tone, instead, we employ data augmentation to obtain tone-disturbed images. Although adversarial attack does work for tone, we found it simpler and more efficient to use data augmentation to deal with it.

GiFD Phase: Gaze-Irrelevant Feature Disturbation

In this part, we will discuss the GiFD Phase (Fig. 3(I)) in detail, which addresses the Prerequisite 1.

This phase aims to disturb the training data from specific gaze-irrelevant factors, such as Identity ($\mathcal{I}$), Expression ($\mathcal{E}$), Illumination ($\mathcal{U}$) and Tone ($\mathcal{T}$), so that they can not be fitted. To achieve this, we use adversarial attack or data augmentation to disturb the training data $\mathbf{x}$ according to the gaze-irrelevant factors.

For Identity ($\mathcal{I}$), Expression ($\mathcal{E}$) and Illumination ($\mathcal{U}$), we use adversarial attack to generate the disturbed data. Without loss of generality, we take Identity ($\mathcal{I}$) for example. We use the Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2014), a fast and commonly-used adversarial attack method, to generate disturbed data:

$$\widetilde{\mathbf{x}} \leftarrow \mathbf{x} - \epsilon \cdot \mathrm{sign}(\frac{\partial \mathcal{L}_{ide}}{\partial \mathbf{x}}), \tag{7}$$

where $\mathbf{x}$ denotes original input image and $\widetilde{\mathbf{x}}$ denotes adversarial sample. Now $\widetilde{\mathbf{x}}$ is the desired data after disturbation, which can not fit to Identity ($\mathcal{I}$) factor. We empirically set $\epsilon = 0.1$ in our experiments. Tab. 2 shows the performance of different ϵ. For Expression ($\mathcal{E}$) and Illumination ($\mathcal{U}$), we use the same strategy to generate disturbed data which can not fit to Expression ($\mathcal{E}$) factor and Illumination ($\mathcal{U}$) factor, respectively.

As shown in Fig. 4(b), the adversarial samples on one specific branch decreases the performance of gaze-irrelevant factor estimation. It can be observed that the corresponding gaze-irrelevant factors have been disturbed.

For Tone ($\mathcal{T}$), we use `ColorJitter` in PyTorch as data augmentation strategy to generate $\widetilde{\mathbf{x}}$. The disturbed data of Tone ($\mathcal{T}$) contains different combinations of tones. They teach the model to learn face images with different combinations of tones with the same gaze label.

Note that we generate the disturbed data for each factor independently.

GcFE Phase: Gaze-Consistent Feature Extraction

In this part, we will describe the GcFE Phase (Fig. 3(II)) details, which solves Prerequisite 2.

The goal of this phase is to guarantee the extracted features $\mathbf{z}$ can fit to gaze factor $\mathcal{G}$ and cannot fit to the gaze-irrelevant factors, such as $\mathcal{I}, \mathcal{E}, \mathcal{U}$ and $\mathcal{T}$, then to make sure the gaze information $\mathcal{G}$ in the dataset is correct and can be learned. To obtain this goal, we feed disturbed data $\widetilde{\mathbf{x}}$ and original data $\mathbf{x}$ into the gaze model to obtain gaze feature $\widetilde{\mathbf{z}}$ and gaze feature $\mathbf{z}$, respectively. Then we minimize the distance between these two features to make sure the gaze estimation model G can extract gaze-consistent features.

We generate $\widetilde{\mathbf{z}}$ and $\mathbf{z}$ by feeding $\widetilde{\mathbf{x}}$ and $\mathbf{x}$ into the gaze

estimation branch F respectively:

$$\widetilde{\mathbf{z}} \leftarrow F(\widetilde{\mathbf{x}}), \mathbf{z} \leftarrow F(\mathbf{x}). \tag{8}$$

To force the gaze estimation model G to extract gaze-consistent features, we minimize the distance between the $\widetilde{\mathbf{z}}$ and $\mathbf{z}$ to align these two features. Now we need to find a proper distance metric. In domain adaption tasks, Maximum Mean Discrepancy (MMD) (Ghifary, Kleijn, and Zhang 2014) is commonly used to measure the distance between the features in source domain and target domain. Moreover, we compare different distance metrics to align features. As shown in Tab. 2, L_{MMD} achieves the best performance. So we use MMD function L_{MMD} described above as feature alignment distance function:

$$\mathcal{L}_{ali} = L_{\text{MMD}}(\mathbf{z}, \widetilde{\mathbf{z}}). \tag{9}$$

Note that if we use more than one factors to generate $\widetilde{\mathbf{z}}$, we will use the mean of such function of every factor for training. For example, if we need to generate $\widetilde{\mathbf{z}}$ on Identity and Expression, let $\widetilde{\mathbf{z}}_{ide}$ be the $\widetilde{\mathbf{z}}$ of Identity and $\widetilde{\mathbf{z}}_{exp}$ be the $\widetilde{\mathbf{z}}$ of Expression, then $\mathcal{L}_{ali} = (L_{\text{MMD}}(\mathbf{z}, \widetilde{\mathbf{z}}_{ide}) + L_{\text{MMD}}(\mathbf{z}, \widetilde{\mathbf{z}}_{exp}))/2$.

Total Loss Function

The total loss function is:

$$\mathcal{L} = \mathcal{L}_{gaze} + \lambda_{ide}\mathcal{L}_{ide} + \lambda_{exp}\mathcal{L}_{exp} + \lambda_{illu}\mathcal{L}_{illu} + \lambda_{ali}\mathcal{L}_{ali}, \tag{10}$$

where $\mathcal{L}_{gaze}$,$\mathcal{L}_{ide}$,$\mathcal{L}_{exp}$,$\mathcal{L}_{illu}$ and $\mathcal{L}_{ali}$ denote the gaze, identity, expression and illumination branch loss function, and the feature alignment distance function, respectively. $\lambda_{ide}, \lambda_{exp}, \lambda_{illu}, \lambda_{ali}$ are hyper-parameters. We empirically set $\lambda_{ide} = \lambda_{exp} = \lambda_{illu} = \lambda_{ali} = 1.0$.

Training and Inference

Algorithm 1 describes the training procedure of our proposed method. During training, only source domain dataset $\mathcal{D}_s$ is needed. For each batch of data, we first generate adversarial examples or augmented examples (GiFD Phase), then optimize the loss function (GcFE Phase).

In inference phase (see Fig.3(III)), we only use gaze branch ($\mathcal{G}$) with target domain dataset $\mathcal{D}_t$, and compute mean angular error between predicted gaze directions and ground-truth gaze directions (Eq. 1) to evaluate domain generalization performance.

Algorithm 1: Training Procedure of Our Proposed Method

Input: Training Dataset on Source Domain $\mathcal{D}_s$
Parameter: $\lambda_{ide}, \lambda_{exp}, \lambda_{illu}, \lambda_{ali}, \epsilon$
Output: $G_\theta(\cdot)$

for $i \leftarrow 1$ *to* N_s **do**
  Randomly sample $(\mathbf{x}_s, \mathbf{y}_s) \leftarrow \mathcal{D}_s$.
  # GiFD Phase
  Calculate $\mathcal{L}_{ide}$ with Eq. (4).
  Generate $\widetilde{\mathbf{z}}_{ide}$ on Identity with Eq. (7) and (8).
  Calculate $\mathcal{L}_{exp}$ with Eq. (5).
  Generate $\widetilde{\mathbf{z}}_{exp}$ on Expression with Eq. (7) and (8).
  Calculate $\mathcal{L}_{illu}$ with Eq. (6).
  Generate $\widetilde{\mathbf{z}}_{illu}$ on Illumination with Eq. (7) and (8).
  Generate $\widetilde{\mathbf{z}}_{tun}$ on Tone using `ColorJitter`.
  # GcFE Phase
  Calculate $\mathcal{L}_{gaze}$ on gaze branch with Eq. (3).
  Calculate $\mathcal{L}_{ide}$, $\mathcal{L}_{exp}$, $\mathcal{L}_{illu}$ on gaze-irrelevant factor branches with Eq. (4), (5) and (6).
  $\mathcal{L}_{ali} \leftarrow \mathbf{z}, \widetilde{\mathbf{z}}_{ide}, \widetilde{\mathbf{z}}_{exp}, \widetilde{\mathbf{z}}_{illu}, \widetilde{\mathbf{z}}_{tun}$ with Eq. (9).
  Train $G_\theta(\cdot)$ and gaze-irrelevant factor branches with Eq. (10).
end for
return $G_\theta(\cdot)$.

Experimental Results

Preparation

Datasets In this paper, we use 4 commonly-used gaze datasets, *i.e.*, ETH-XGaze ($\mathcal{D}_E$) (Zhang et al. 2020), Gaze360 ($\mathcal{D}_G$) (Kellnhofer et al. 2019), MPIIGaze ($\mathcal{D}_M$) (Zhang et al. 2017a) and EyeDiap ($\mathcal{D}_D$) (Funes Mora, Monay, and Odobez 2014).

Since ETH-XGaze and Gaze360 have larger gaze distributions than MPIIGaze and EyeDiap (Liu et al. 2021), we use ETH-XGaze ($\mathcal{D}_E$) and Gaze360 ($\mathcal{D}_G$) datasets for training, and use MPIIGaze ($\mathcal{D}_M$) and EyeDiap ($\mathcal{D}_D$) for evaluation. As a result, we perform 4 domain generalization tasks, *i.e.*, $\mathcal{D}_E \rightarrow \mathcal{D}_M$, $\mathcal{D}_E \rightarrow \mathcal{D}_D$, $\mathcal{D}_G \rightarrow \mathcal{D}_M$ and $\mathcal{D}_G \rightarrow \mathcal{D}_D$.

ETH-XGaze contains 80 subjects, and we use data from 75 subjects for training (713646 images) and the rest 5 subjects data for validation. For Gaze360, we only use the front-side face images for training, the total number is 84902. For these two datasets, we use the pre-processed data provided by the authors. For MPIIGaze, we follow the method in (Sugano, Matsushita, and Sato 2014) to rectify the dataset and use the full dataset with 45000 images for evaluation. For EyeDiap, we follow (Zhang et al. 2017a) to sample the images for every 15 frames in the VGA videos of screen targets session. The processed dataset contains 16674 images.

Comparison Methods For Baseline, we use the same model as our proposed method, and only train the gaze estimation branch using L_1 loss between the predicted gaze direction and its ground truth. For typical gaze estimation methods, we use RT-Gene (Fischer, Chang, and Demiris 2018), Dilated-Net (Chen and Shi 2018), Full-Face (Zhang et al. 2017a) and CA-Net (Cheng et al. 2020a) for comparison and use the results reported in (Cheng, Bao, and Lu 2022) for reference. For domain generalization methods, to the best of our knowledge, PureGaze (Cheng, Bao, and Lu 2022) is the first and the only domain generalization method on gaze estimation. Therefore, we choose the PureGaze for comparison and use the results reported by the author. For reference, we also use SOTA unsupervised domain adaption methods, including ADDA (Tzeng et al. 2017), DAGEN (Guo et al. 2020), GazeAdv (Wang et al. 2019), Gaze360 (Kellnhofer et al. 2019), PnP-GA (Liu et al. 2021), RUDA (Bao et al. 2022) and CRGA (Wang et al. 2022) for comparison. Note that these methods usually require a small amount of unlabeled target domain data.

Factor	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$	Avg
Baseline	8.80	8.52	8.18	9.00	8.63
Identity ($\mathcal{I}$)	7.92	7.49	8.10	9.73	8.31
Expression ($\mathcal{E}$)	<u>6.85</u>	7.14	7.81	8.79	7.65
Illumination ($\mathcal{U}$)	7.29	**6.52**	7.73	8.53	<u>7.52</u>
Tone ($\mathcal{T}$)	7.62	8.61	7.68	9.10	8.25
$\mathcal{I}+\mathcal{E}$	7.84	7.42	8.62	9.07	8.24
$\mathcal{I}+\mathcal{U}$	7.50	<u>6.91</u>	8.35	8.64	7.85
$\mathcal{I}+\mathcal{T}$	7.42	6.97	7.76	**7.76**	**7.48**
$\mathcal{E}+\mathcal{U}$	7.62	7.37	8.47	9.01	8.12
$\mathcal{E}+\mathcal{T}$	<u>6.85</u>	7.38	7.60	9.34	7.67
$\mathcal{U}+\mathcal{T}$	6.89	8.06	**7.44**	8.93	7.83
$\mathcal{I}+\mathcal{E}+\mathcal{U}$	6.89	7.78	7.82	8.52	7.75
$\mathcal{I}+\mathcal{E}+\mathcal{T}$	7.48	7.17	7.59	8.59	7.71
$\mathcal{I}+\mathcal{U}+\mathcal{T}$	7.47	7.31	7.82	<u>8.19</u>	7.70
$\mathcal{E}+\mathcal{U}+\mathcal{T}$	7.01	7.34	7.56	8.91	7.71
$\mathcal{I}+\mathcal{E}+\mathcal{U}+\mathcal{T}$	**6.50**	7.44	<u>7.55</u>	9.03	7.63

Table 1: Results of four domain generalization tasks when using different combinations of gaze-irrelevant factors to generate disturbed data. Angular error in degrees are shown. Bold and underline denote the best and the second best result among each column, respectively.

Implementation Details We use a single NVIDIA GPU to run the experiments. ResNet-18 (He et al. 2016) is used as backbone for all the experiments. All the images are resized to 224×224 and normalized to $[0, 1]$. The batch size is 128, and we train the model for 10 epochs for ETH-XGaze and 100 epochs for Gaze360. We use Adam optimizer with a learning rate of 10^{-4}, and set $\beta_1 = 0.9$, $\beta_2 = 0.95$.

Gaze-Irrelevant Feature Disturbation Results

In this part, we will try different combinations of gaze-irrelevant factors to generate disturbed data, then evaluate the performance on four domain generalization tasks, in order to know which combination is the most effective.

Tab. 1 shows the results. It can be seen that in most cases, handling different combinations of gaze-irrelevant factors improves domain generalization performance over Baseline. Disturbing Identity and Tone together ($\mathcal{I}+\mathcal{T}$) achieves the best average domain generalization performance on the four tasks, and disturbing Illumination only ($\mathcal{I}$) results in the second best average performance. Disturbing all four factors ($\mathcal{I}+\mathcal{E}+\mathcal{U}+\mathcal{T}$) achieves the third best performance. and we use this combination for latter discussion.

To obtain the optimal model parameters, we compare model performances of under different ϵ and distance metrics. Tab. 2 shows the results. We observed that $\epsilon = 0.1$ achieves the best performance on $\mathcal{D}_E \rightarrow \mathcal{D}_M$, $\mathcal{D}_E \rightarrow \mathcal{D}_D$ and $\mathcal{D}_G \rightarrow \mathcal{D}_D$ tasks, and achieves a close-to-best performance on the $\mathcal{D}_G \rightarrow \mathcal{D}_M$ task. For distance metrics, we compare four different distance metrics, *i.e.*, L_1, L_2, Jensen-Shannon (JS) divergence (L_{JS}) and MMD (L_{MMD}). It can be observed that MMD (L_{MMD}) achieves the best performance on $\mathcal{D}_E \rightarrow \mathcal{D}_M$ and $\mathcal{D}_G \rightarrow \mathcal{D}_M$ tasks, and achieves the best average performance on four tasks. Thus, we set the intensity of adversarial perturbation $\epsilon = 0.1$, and choose MMD as the distance metric.

	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$	Avg
$\epsilon = 0.01$	7.11	7.85	7.64	9.15	7.94
$\epsilon = 0.05$	6.87	7.89	**7.50**	9.45	7.93
$\epsilon = 0.1$	**6.50**	**7.44**	7.55	**9.03**	**7.63**
$\epsilon = 0.2$	7.02	8.16	11.55	12.17	9.73
L_1	7.29	**6.89**	8.00	9.56	7.94
L_2	7.23	7.70	7.66	8.35	7.74
L_{JS}	6.95	7.82	8.85	**8.28**	7.98
L_{MMD}	**6.50**	7.44	**7.55**	9.03	**7.63**
w/o $\mathcal{L}_{ali}$	7.48	8.15	7.92	9.66	8.30
w/ $\mathcal{L}_{ali}$	**6.50**	**7.44**	**7.55**	**9.03**	**7.63**

Table 2: Cross-dataset validation results of different adversarial noise intensity ϵ, different distance metrics of feature and w/ or w/o $\mathcal{L}_{ali}$. Results show the angular error in degrees. "w/o $\mathcal{L}_{ali}$" indicates directly using the disturbed data to train the model, without feature alignment terms.

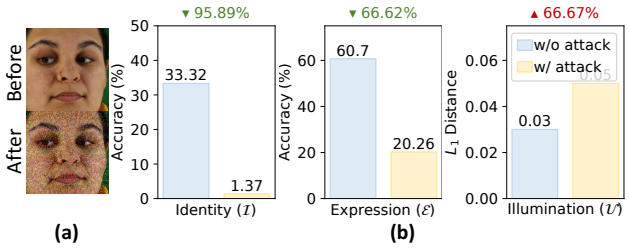


Figure 4: (a) Samples before/after adversarial attack. (b) Adversarial attack results of different gaze-irrelevant factors.

Moreover, to further show the necessity of Gaze-consistent Feature Extraction Phase, *i.e.*, feature alignment terms $\mathcal{L}_{ali}$, we remove the feature alignment terms $\mathcal{L}_{ali}$ and train the model using disturbed data and original data directly. Results are shown in the last row in Tab. 2, which indicate the performance drop when removing the feature alignment terms and show the necessity of Gaze-consistent Feature Extraction Phase.

Furthermore, our proposed method use adversarial attack to generate disturbed data on Identity, Expression and Illumination. We randomly choose 10000 samples from the training set of $\mathcal{D}_E$, called $\mathcal{D}_{adv}$ and perform adversarial attack on the branch of one of these factors. For Identity and Expression, we evaluate the classification accuracy (the higher the better) before and after adversarial attack, respectively. For Illumination, we evaluate L_1 distance (the lower the better) between $\widetilde{\mathbf{y}}_{illu}$ and $\mathbf{y}_{illu}$ on $\mathcal{D}_{adv}$ before and after adversarial attack. Samples before and after adversarial attack are shown in Fig. 4(a). Results in Fig. 4(b) show that adversarial attack is effective to lower the performance on specific tasks of gaze-irrelevant factors, and it has the potential to generate the disturbed data that can not fit to gaze-irrelevant factors.

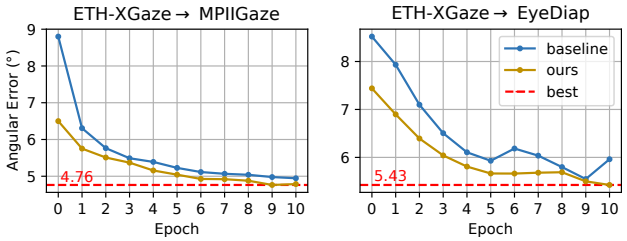


Figure 5: Fine-tuning with 100 labeled samples target domain using Baseline and Ours. Dashed line indicates the best results, which are all achieved by our proposed method.

Comparison with SOTA Methods

In this part, we first compare our proposed method with existing domain generalization (DG) methods and typical gaze estimation methods to evaluate the effectiveness of our proposed method on domain generalization task. Note that this task forbids to use any target domain information. Results are shown in the second row of Tab. 3. It can be observed that our proposed method ($\mathcal{I}+\mathcal{E}+\mathcal{U}+\mathcal{C}$) surpasses all the SOTA methods in domain generalization task. For example, our method outperforms the Baseline and PureGaze by 1° and 0.66° in average, respectively, whose performance improvement is significant. Meanwhile, our proposed method surpasses the Baseline on three domain generalization tasks, and achieves similar performance for $\mathcal{D}_G \rightarrow \mathcal{D}_D$. These results indicate the effectiveness of our proposed method on domain generalization task.

In addition, we compare our proposed method with unsupervised domain adaption (UDA) tasks on gaze estimation, as shown in the third row of Tab. 3. Note that unsupervised domain adaption methods require a small number of unlabeled target domain samples. It can be seen that our proposed domain generalization method with no access to target domain information can even surpass most of unsupervised domain adaption methods on task $\mathcal{D}_E \rightarrow \mathcal{D}_D$, which demonstrates the strength of our proposed method as it does not need any target domain information.

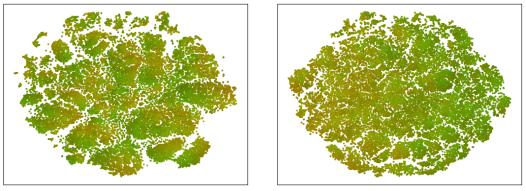
To further demonstrate the effectiveness of the proposed framework, we fine-tune the baseline model and our proposed model, *i.e.*, further train the model trained on source domain, on 100 randomly chosen target domain samples with labels for 10 epochs, results on supervised domain adaption task are shown in the last row of Tab. 3. These results show that our proposed method always achieves better results after fine-tuning compared with fine-tuned baseline. Furthermore, we show results of each epoch in Fig. 5 for task $\mathcal{D}_E \rightarrow \mathcal{D}_M$ and $\mathcal{D}_E \rightarrow \mathcal{D}_D$. It can be observed that for most of time, the fine-tuned proposed method surpasses the fine-tuned baseline. This result further proves that the proposed method is easier to fine-tune on the target domain.

Analysis of Gaze-Consistent Feature Extraction

We further analyze the extraction of gaze-consistent feature. Gaze-consistent feature is extracted from original input images $\mathbf{x}$ and is expected to be more correlated to gaze and less correlated to gaze-irrelevant factors. To test this hypothesis, we take $\mathcal{D}_E \rightarrow \mathcal{D}_M$ task as an example, and use t-SNE (Van der Maaten and Hinton 2008) to visualize gaze feature vectors extracted by Baseline and our proposed method. Fig. 6 shows the visualization results, where similar colors represents similar gaze directions. We observe that the gaze feature extracted by the Baseline is irrelevant to gaze, as features of similar gaze directions are far apart (see Fig. 6(a)). Interestingly, the gaze feature extracted by our proposed method is highly correlated to gaze, as features of similar gaze directions are close (see Fig. 6(b)). From these experimental results, it can be concluded that our proposed method is able to extract gaze-consistent feature.

Task	Methods	$\lvert\mathcal{D}_t\rvert$	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
DG	Baseline	0	8.80	8.52	8.18	**9.00**
	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
	PureGaze	0	7.08	7.48	9.28	9.32
	Ours	0	**6.50**	**7.44**	**7.55**	9.03
UDA	ADDA	500	5.77	11.12	7.18	12.56
	GazeAdv	100	7.26	8.37	7.88	9.81
	Gaze360	100	6.23	7.80	7.00	8.77
	DAGEN	500	5.68	7.92	8.02	11.08
	PnP-GA	10	5.53	5.87	6.18	7.92
	RUDA	100	5.70	7.52	6.20	7.02
	CRGA	>0	5.48	5.66	5.89	6.49
SDA	Baseline*	100	4.95	5.55	5.41	5.71
	Ours*	100	**4.76**	**5.43**	**5.34**	**5.66**

Table 3: Results of SOTA methods. Angular error in degrees are shown. Bold denotes the best result among each column on one specific task. * denotes the model is fine-tuned using 100 randomly-chosen labeled target domain samples.



(a) Baseline: $\mathcal{D}_E \rightarrow \mathcal{D}_M$ (b) Ours: $\mathcal{D}_E \rightarrow \mathcal{D}_M$

Figure 6: t-SNE visualization results of gaze feature. Note that similar colors represent similar gaze directions.

Conclusion

In this paper, we propose a novel gaze domain generalization method based on gaze-consistent features. We specifically disturb training data against gaze-irrelevant factors using adversarial attack (for Identity, Expression and Illumination) or data augmentation (for Tone), then align the gaze features from disturbed data with non-disturbed gaze features using MMD. Our proposed method achieves state-of-the-art performance on domain generalization task. We find that all the proposed gaze-irrelevant factors can be disturbed and the gaze features extracted by our method are more correlated to gaze, which improves the explainability of our approach.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) under Grant 61972012.

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 (ICPR)*, 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. E. 2018. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, 309–324. Springer.

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.; Lu, F.; and Zhang, X. 2018. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 100–115.

Cheng, Y.; Wang, H.; Bao, Y.; and Lu, F. 2021. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*.

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.

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

Funes Mora, K. A.; Monay, F.; and Odobez, J.-M. 2014. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*, 255–258.

Ghifary, M.; Kleijn, W. B.; and Zhang, M. 2014. Domain adaptive neural networks for object recognition. In *Pacific Rim international conference on artificial intelligence*, 898–904. Springer.

Goodfellow, I. J.; Erhan, D.; Carrier, P. L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, 117–124. Springer.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. arXiv:1412.6572.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, D.; Song, L.; and Zhi, C. 2010. Multi-illumination face recognition from a single training image per person with sparse representation. In *Asian Conference on Computer Vision*, 672–682. Springer.

Ilse, M.; Tomczak, J. M.; Louizos, C.; and Welling, M. 2020. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, 322–348. PMLR.

Katsini, C.; Abdrabou, Y.; Raptis, G. E.; Khamis, M.; and Alt, F. 2020. The Role of Eye Gaze in Security and Privacy Applications: Survey and Future HCI Research Directions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367080.

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.

Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; and Tao, D. 2018. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 624–639.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

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 *2011 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 *5th Berkeley Symp. Math. Statist. Probability*, 281–297.

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

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556.

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.

Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; and Abbeel, P. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 23–30. IEEE.

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.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, H.; Dong, X.; Chen, Z.; and Shi, B. E. 2015. Hybrid gaze/EEG brain computer interface for robot arm control on a pick and place task. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1476–1479.

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.; Zhao, Y.; and Lu, F. 2022. Control with Vergence Eye Movement in Augmented Reality See-Through Vision. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 548–549.

Xu, P.; Sugano, Y.; and Bulling, A. 2016. Spatio-temporal modeling and prediction of visual attention in graphical user interfaces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3299–3310.

Yi, D.; Lei, Z.; Liao, S.; and Li, S. Z. 2014. Learning face representation from scratch. arXiv:1411.7923.

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 *European Conference on Computer Vision*, 365–381. Springer.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2015. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4511–4520.

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 (CVPR) Workshops*.

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.

Zhang, Y.; Wang, H.; and Shi, B. E. 2021. Gaze-controlled Robot-assisted Painting in Virtual Reality for Upper-limb Rehabilitation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 4513–4517. IEEE.