From Feature to Gaze: A Generalizable Replacement of Linear Layer for Gaze Estimation

Yiwei Bao Feng Lu *
State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University
{baoyiwei, lufeng}@buaa.edu.cn

Abstract

Deep-learning-based gaze estimation approaches often suffer from notable performance degradation in unseen target domains. One of the primary reasons is that the Fully Connected layer is highly prone to overfitting when mapping the high-dimensional image feature to 3D gaze. In this paper, we propose Analytical Gaze Generalization framework (AGG) to improve the generalization ability of gaze estimation models without touching target domain data. The AGG consists of two modules, the Geodesic Projection Module (GPM) and the Sphere-Oriented Training (SOT). GPM is a generalizable replacement of FC layer, which projects high-dimensional image features to 3D space analytically to extract the principle components of gaze. Then, we propose Sphere-Oriented Training (SOT) to incorporate the GPM into the training process and further improve cross-domain performances. Experimental results demonstrate that the AGG effectively alleviate the overfitting problem and consistently improves the cross-domain gaze estimation accuracy in 12 cross-domain settings, without requiring any target domain data. The insight from the Analytical Gaze Generalization framework has the potential to benefit other regression tasks with physical meanings.

1. Introduction

Eye gaze reveals where human attention lands, which has been widely applied in a variety of territories, such as VR/AR systems [3, 15, 26], medical analysis [4, 13, 14] and human-computer interaction[17, 27, 34]. Gaze estimation methods can be classified into two categories: model-based approaches and appearance-based approaches. Both approaches have their own strengths and weaknesses. Model-based approaches estimate gaze by modeling the anatomical structure of the eyeball. These methods achieve remarkable accuracy in controlled environment. But they typically require dedicated hardware such as infrared cameras and light sources. Appearance-based approaches use cost-effective web cameras. They typically train Convolutional Neural Networks (CNNs) in an end-to-end way, enabling them to predict gaze direction from user face/eye images directly. In recent years, appearance-based approaches have garnered great interest due to their simplified hardware requirements and the potential for widespread applications.

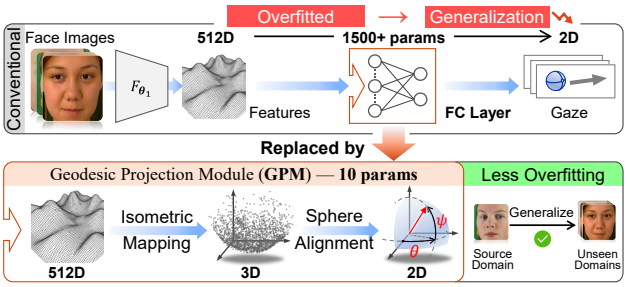

Figure 1. Overview of the proposed AGG framework for generalizing gaze estimation models to unseen target domains.

However, appearance-based methods suffer from severe performance degradation in cross-domain settings. To improve the cross-domain performance, various domain adaptation approaches have been proposed, *i.e.*, adversarial learning [12, 32], contrastive learning [33] and collaborative learning [20]. However, these approaches require a number of target domain samples for adaptation, which is not always attainable in real-world scenarios. More recently, Cheng *et al*. proposed to generalize gaze estimation model by purifying gaze feature during source domain training [8]. The gaze generalization task is more practical yet more challenging, because it does not have access to any target domain data.

One of the significant reasons for the poor cross-domain performances is the overfitting problem. Gaze estimation CNNs are trained to extract high-dimensional image features (*e.g.* 512D) from input face images and map these features to gaze (3D unit vector) by a Fully Connected layer with thousands of parameters. **The numerous parameters of the FC layer easily overfit to gaze irrelevant factors within the high-dimensional image features during the**

*Corresponding Author. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62372019.

end-to-end training process. One possible solution is to extract gaze-related information from the image features, *i.e*. the Principle Component of Gaze (PCG), while excluding other irrelevant information.

In this paper, we introduce the Analytical Gaze Generalization framework(AGG), a novel gaze generalization approach that connects the high-dimensional image features to gaze analytically. The AGG consists of two modules, the Geodesic Projection Module (GPM) and the Sphere-Oriented Training (SOT) module. Given a pretrained gaze estimation model, **the GPM serves as a replacement of the last FC layer, alleviating the overfitting issue by analytical projection and alignment.** Based on the observation that the geodesic distance between the image features is proportional to the angular gaze difference between samples, the GPM projects the high-dimensional image features to 3D space by the geodesic distance to extract the Principle Component of Gaze. Then, we estimate gaze from the projected features by the proposed Sphere Alignment algorithm using physical rotation and scaling with only 10 learnable parameters. Next, we propose the Sphere-Oriented Training to incorporate the GPM into the training process to improve the generalization ability of the whole network. Given source domain labels, the Sphere-Oriented Training optimizes the gaze estimation network based on the reverse process of the GPM.

Experiments show that, **by only replacing the last FC layer of the baseline model with GPM, both the accuracy and the stability in cross-domain testing have been improved.** After optimizing the model using the Sphere-Oriented Training in the source domain, the performance is further improved and outperforms SOTA gaze estimation methods in multiple different cross-domain settings. The primary contributions of this work are as follow:

- We propose the Geodesic Projection Module (GPM), a novel method that predicts gaze from the geodesic distance between the image features analytically. As a novel and explainable approach for gaze estimation, the insight GPM presents may also inspire other regression tasks like pose estimation.
- We propose the AGG framework for generalizable gaze estimation. The AGG framework utilizes the Sphere-Oriented Training module to optimize the gaze estimation model based on the reverse process of the GPM for better generalization ability.
- Experimental results illustrate that the proposed AGG achieves consistent improvements in 12 different cross-dataset settings. The AGG improves the generalization ability of the baseline model up to 35.79% without touching target domain data.

The subsequent sections are organized as follow: in Sec. 3, we present the motivation and design principles of the AGG through several validation experiments. In Sec. 4, we provide a detailed introduction to the AGG framework. In Sec. 5, we assess the proposed method both quantitatively and qualitatively.

2. Related Work

2.1. Gaze Estimation

There are two mainstream gaze estimation approaches, the model-based approaches and the appearance-based approaches. Model-based approaches estimate gaze by reconstructing the anatomy structure of the eyeball [11]. These methods achieve remarkable accuracy but also require personal calibration and dedicated devices such as depth sensors [28, 35], infrared cameras [10, 29] and lights [10, 19].

Appearance-based approaches usually estimate gaze from user images captured by a single web camera. Early methods estimate gaze from eye images by traditional machine learning algorithms like manifold embedding [25] and adaptive linear regression [21]. Lu *et al*. propose to estimate eye rotation by measuring the geodesic distance between eye images [22]. Wang *et al*. propose to combine the eye appearance with eye geometry by a Hierarchical Generative Model []. More recently, a number of gaze estimation datasets have been collected [9, 12, 16, 38, 39]. These datasets provide hundreds of thousands of user images with gaze labels, which makes deep-learning-based gaze estimation possible. Representative studies include gaze estimation using convolutional neural networks (CNNs) [37] with eye images [7, 37] or face images [1, 5, 6, 16, 38]. Some previous studies also represent gaze features as low dimensional manifolds for personalization [23] and unsupervised learning [36]. But these methods still construct manifolds by data-driven learning approach with supervision like gaze redirection. Our method analytically connects the high-dimensional image feature to gaze.

2.2. Cross-domain Gaze Estimation

One of the major problem of the deep-learning-based approaches is that the performance degrades severely when testing on a different domain. To improve the cross-domain performance, a number of unsupervised domain adaptation methods have been proposed. Liu *et al*. propose to adapt the model to target domain with the guidance of outliers by collaborative learning [20]. Wang *et al*. utilize contrastive learning to pull features with close gaze labels together [33]. Bao *et al*. propose to improve the cross-dataset accuracy by the rotation consistency of gaze [2]. Nevertheless, above methods require target domain images to train domain specific models, which is infeasible in real world settings, as target domain data is often inaccessible. Recently, Cheng *et al*. propose to improve the generalization ability of gaze estimation model by purifying gaze feature in source domain[8]. The gaze generalization problem without

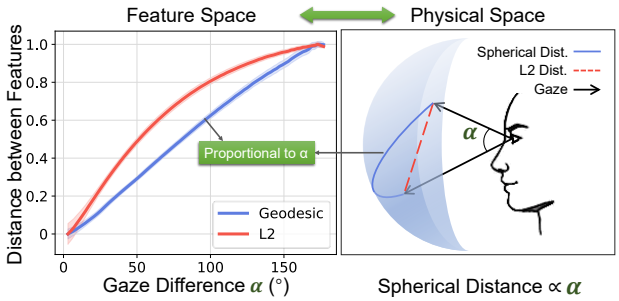

Figure 2. Observation: The gaze differences between samples can be linearly represented by the geodesic distances between extracted features. Such characteristic of the feature space is consistent with the physics of gaze: the distance along the spherical surface is proportional to the gaze differences.

access to target domain data is more challenging and yet to be solved.

3. Analytical Gaze Estimation from Features

The aim of this paper is to design a replacement of the last regression FC layer to alleviate the overfitting problem. In this section, we try to answer the following key questions by a series of validation experiments:

Question 1: How many dimensions does the Principle Components of Gaze (PCG) in the high-dimensional image features have?

Question 2: How to extract the PCG from the high-dimensional image features for generalizable gaze estimation?

3.1. The Overfitting Problem

To explore above questions, we initially pretrain a commonly used baseline model, *i.e.* ResNet-18 in the source domain (ETH-XGaze [39]) for analysis.

Given the source domain $\mathcal{D}_s = \{\boldsymbol{x}_i, \boldsymbol{y}_i|_{i=1}^N\}$ where $\boldsymbol{x}_i$ is the face image and $\boldsymbol{y}_i = (x_i, y_i, z_i)$ is the ground truth unit gaze direction vector, we pretrain the gaze estimation model by $\mathcal{L}_1$ loss function:

$$\begin{aligned} \boldsymbol{f}_i &= F_{\boldsymbol{\theta}_1}(\boldsymbol{x}_i), \\ \arg\min_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2} &(\mathcal{L}_1(\boldsymbol{y}_i, L_{\boldsymbol{\theta}_2}(\boldsymbol{f}_i))|_{i=1}^N). \end{aligned} \tag{1}$$

where $F_{\boldsymbol{\theta}_1}(\cdot)$ is the feature extractor CNN, $\boldsymbol{f}_i$ is the 512D high-dimensional image feature and $L_{\boldsymbol{\theta}_2}(\cdot)$ is the last Fully Connected layer which estimates gaze from $\boldsymbol{f}_i$.

Previous study [8] has proven that the feature extracted by the pretraind baseline model $\boldsymbol{f}_i$ encompasses not only gaze information, but also other visual contents, including appearance, illumination and head pose. Thus, the dimension of the PCG should be less than the dimension of $\boldsymbol{f}_i$ itself. Combined with the fact that gaze direction is a 3D unit vector with 2 degrees of freedom, **for the answer of Question 1, we hypothesis that the theoretical minimum dimension of the PCG should be 2D, identical to the gaze ground truth.** The regression FC layer $L_{\boldsymbol{\theta}_2}$ is at a high risk of overffiting since the number of parameters in $L_{\boldsymbol{\theta}_2}$ and the dimension of $\boldsymbol{f}_i$ are way beyond the minimum required quantity.

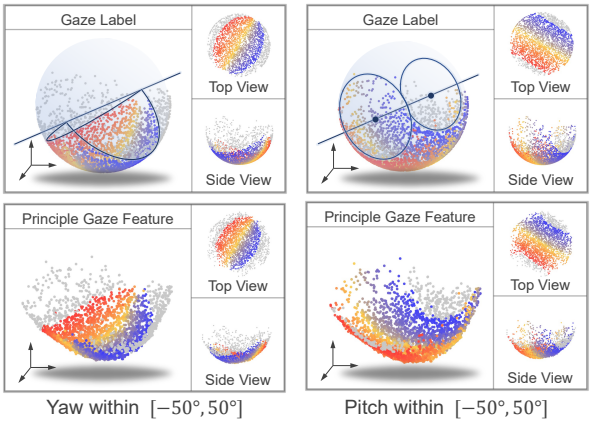

Figure 3. Following the observation in Fig. 2, we construct PGF by projecting high-dimensional features to the 3D space using geodesic distance. The PGF shares the same spherical distribution pattern as the gaze label. Data points are colored by gaze yaw and pitch angles respectively.

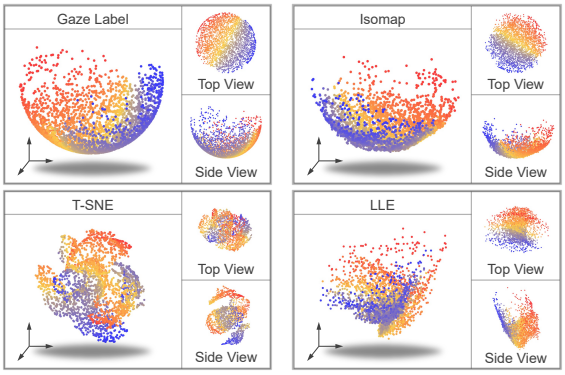

Figure 4. Projecting features extracted by the pretrain gaze estimation model into 3D space by varies distance metrics. Projections by geodesic distance (Isomap) show identical spherical distribution pattern as the gaze label. Colored by gaze yaw angles.

3.2. Gaze on the 3D Sphere

To answer **Question 2**, we examine the distribution of the gaze ground truth. In the physical space, gaze distributes across the surface of the 3D unit sphere. The distance alone the data manifold *i.e.* the spherical surface is proportional to the angular differences between gaze directions.

To verify if similar relationship also exists in the feature space, we visualize the distance alone the data manifold in

the feature space, *i.e*. the geodesic distance between $\boldsymbol{f}_i$ in Fig. 2. The result reveals an important observation:

Observation: the geodesic distance between $\boldsymbol{f}_i$ is in a strong direct proportion to the angular gaze differences between samples.

This observation leads to **a possible answer of Question 2: The PCG within the high-dimensional image features can be extracted by the geodesic distance.** In the following step, we leverage the geodesic distance to extract the Principle Gaze Feature (PGF) from $\boldsymbol{f}_i$ for generalizable gaze estimation.

3.3. Mapping Features to 3D Space Analytically

According to above observations, we utilize the geodesic distance to extract the PGF from the high-dimensional image features. Specifically, we project the high-dimensional image features to the 3D space using geodesic distance, *i.e*. Isometric Mapping (Isomap) [30]. Results in Fig. 3 demonstrate that the image features after projection (the PGF) share similar distribution pattern as gaze: the PGF distributes across the surface of a 3D sphere approximately. To better demonstrate the distribution pattern, we color the PGFs within certain gaze range with pitch and yaw angles respectively. It is obvious that gaze pitch and yaw angle change monotonically alone the longitudinal and latitudinal direction of the PGF Sphere. The Principle Gaze Feature preserves gaze information while excluding unnecessary factors from $\boldsymbol{f}_i$ to alleviate the overffiting problem. Generalizable gaze estimations could be made by simply aligning the PGF Sphere with the unit sphere of gaze distribution using the gaze label. This alignment process consists of simple physical operation including rotation and scaling with only 10 parameters, which is less unlikely to overfit. We further utilize this idea to optimize the feature extractor CNN $F_{\boldsymbol{\theta}_1}$ for better generalization ability. The detailed implementation will be introduced in Sec. 4.

3.4. Choice of Distance Metrics

In this section, we validate some other possible answers of **Question 2**. Specifically, we project the image features $\boldsymbol{f}_i$ into the 3D space using two other dimension reduction methods: the Local Linear Embedding (LLE) [24] and the T-distributed Stochastic Neighbor Embedding (t-SNE) [31]. It is obvious in Fig. 4 that only the results of Isomap exhibit similar distribution pattern to the gaze label. Although the distributions of other dimension reduction methods also exhibit some directional characteristics, their overall distributions do not show a clear geometric pattern like Isomap. Above results prove that geodesic distance is the key to extract the Principle Components of Gaze from the high-dimensional image feature. In the next section, we explain the specific implementation of the AGG framework

4. Method

We propose the Analytical Gaze Generalization framework, a domain generalization method for gaze estimation. The AGG framework comprises two key modules: the Geodesic Projection Module (GPM) module and the Sphere-Oriented Training (SOT) module. The Geodesic Projection Module predicts gaze analytically from the pretrained image feature by constructing the Principle Gaze Feature using geodesic distance. Next, the Sphere-Oriented Training module optimizes the pretrained gaze estimation network according to the reverse process of the GPM for better generalization ability. The overview of the Analytical Gaze Generalization framework is presented in Fig. 5.

4.1. Geodesic Projection Module

The Geodesic Projection Module mainly consists of two steps. First, we project the image features extracted by the pretrained gaze estimation model into 3D space using geodesic distance, *i.e*. Isomap. The projected feature (named the Principle Gaze Feature) distributes alone the surface of a sphere (named the PGF Sphere) approximately. Then, we align the PGF Sphere with the unit sphere represents the gaze label distribution to predict gaze directions analytically.

First, we pretrain a gaze estimation model consists of a feature extractor CNN $F_{\boldsymbol{\theta}_1}$ and a Fully Connected layer $L_{\boldsymbol{\theta}_2}$ for gaze regression in the source domain $\mathcal{D}_s = \{\boldsymbol{x}_i, \boldsymbol{y}_i|_{i=1}^{N}\}$ using $\mathcal{L}_1$ loss function according to Eq. (1). Then, we freeze the gaze estimation model and extract the high-dimensional image features (512D in our experiments) by $\{\boldsymbol{f}_i = F_{\boldsymbol{\theta}_1}(\boldsymbol{x}_i)|_{i=1}^{N'}\}$. The Principle Gaze Feature $\boldsymbol{e}_i \in \mathbb{R}^3$ is constructed by projecting the image features $\boldsymbol{f}_i$ into the 3D space using Isomap algorithm:

$$\{\boldsymbol{e}_i|_{i=1}^{N'}\} = Isomap(\{\boldsymbol{f}_i|_{i=1}^{N'}\}). \tag{2}$$

Since the Principle Gaze Feature distributes across the surface of a 3D sphere approximately, the next step is to predict gaze directions by aligning this PGF Sphere with the unit gaze label sphere. First, we locate the center of the PGF Sphere $\boldsymbol{O}_c$ and rotate it to align the orientation with the unit gaze label sphere:

$$\boldsymbol{e}_i' = \boldsymbol{R}(\boldsymbol{e}_i - \boldsymbol{O}_c) = (x_i^{e'}, y_i^{e'}, z_i^{e'})^T, \tag{3}$$

where $\boldsymbol{R}$ is the rotation matrix. Then, we calculate the Euler angles of $\boldsymbol{e}_i'$ and predict gaze directions $\boldsymbol{y}_i = (\theta', \psi')$ by simple linear fittings:

$$\begin{cases} \theta_i' = k_1 arctan(\frac{x_i^{e'}}{z_i^{e'}}) + b_1, \\ \psi_i' = k_2 arcsin(y_i^{e'}) + b_2. \end{cases} \tag{4}$$

The final gaze prediction $\boldsymbol{y}_i' = (x_i', y_i', z_i')$ is obtained through converting the Euler angle predictions $(\theta_i', \psi_i', 0)$ to unit direction vectors. We formalize the process

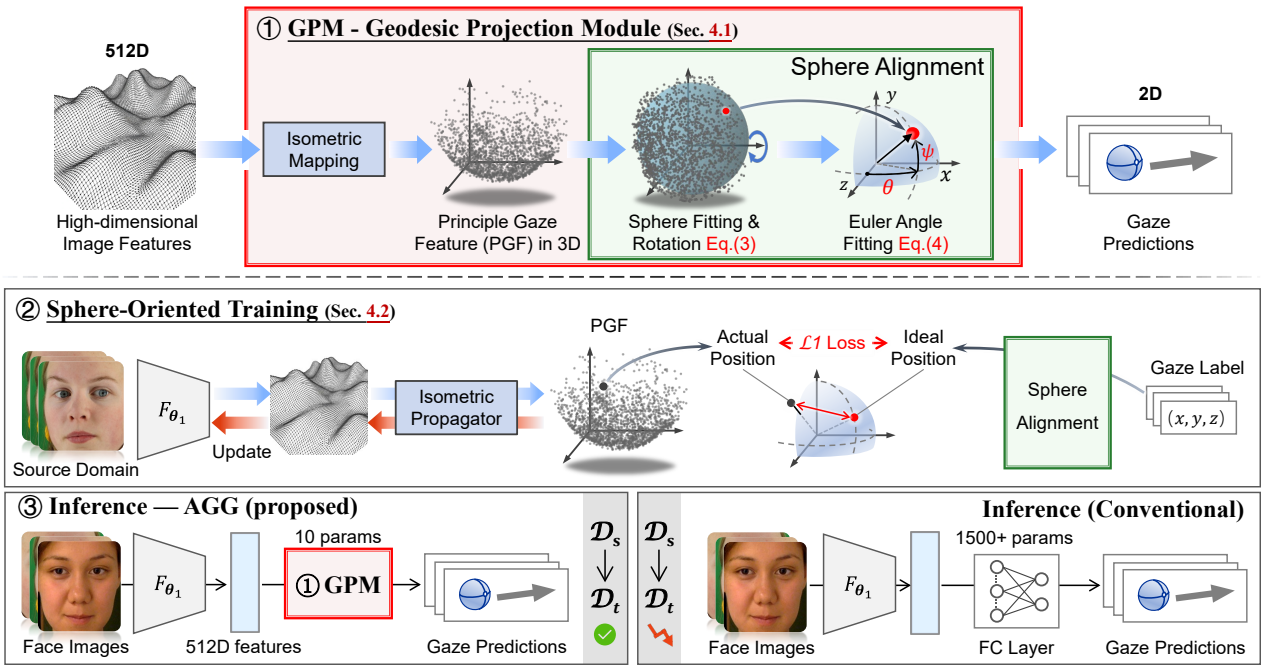

Figure 5. Overview of the proposed AGG framework. We propose two modules, the Geodesic Projection Module (GPM) and the Sphere-Oriented Training module. We replace the last Linear layer of the pretrained baseline model with GPM to estimate gaze from the high-dimensional image feature analytically. The Sphere-Oriented Training optimizes the gaze estimation model according to the reverse process of the GPM for better generalization ability.

from the Principle Gaze Feature $\boldsymbol{e}_i$ to gaze prediction $\boldsymbol{y}'_i$ as the Sphere Alignment algorithm: $\boldsymbol{y}'_i = SA_{\boldsymbol{\theta}_s}(\boldsymbol{e}_i)$, where $\boldsymbol{\theta}_s$ is the set of 10 learnable parameters $\boldsymbol{\theta}_s = \{\boldsymbol{O}_c, \boldsymbol{R}, k_1, k_2, b_1, b_2\}$. These parameters are obtained by minimizing the angular difference between gaze prediction $\boldsymbol{y}'_i$ and the source domian gaze label $\boldsymbol{y}_i$:

$$\arg\min_{\boldsymbol{\theta}_s}(\text{Angular}(\boldsymbol{y}_i, SA_{\boldsymbol{\theta}_s}(\boldsymbol{e}_i))|_{i=1}^{N}). \tag{5}$$

Since the GPM only contains 10 learnable parameters, we only randomly choose 2000 source domain samples to optimize $\boldsymbol{\theta}_s$ in our experiments. In the test time, features of target domain samples are concatenated to the geodesic distance map built in the source domain for Isometric Mapping. The parameters of the SA algorithm remain fixed in the Sphere-Oriented Training and test time.

4.2. Sphere-Oriented Training

The Geodesic Projection Module predicts gaze from the image feature analytically, thus the performance will be affected by the quality of the image feature. To extract generalizable image features, we propose Sphere-Oriented Training to optimize the pretrained feature extractor CNN $F_{\boldsymbol{\theta}_1}$ with the reverse process of the GPM in the source domain.

Given a source domain sample $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$, we could reversely calculate the ideal position of the corresponding Principle Gaze Feature since the Sphere Alignment algorithm is totally analytical: $\hat{\boldsymbol{e}}_i = SA^{-1}(\boldsymbol{y}_i)$. Theoretically, the pretrained CNN could be optimized by minimizing the distance between the ideal position and actual position of the Principle Gaze Feature on the PGF Sphere:

$$\arg\min_{\boldsymbol{\theta}_1}(\mathcal{L}_1(\hat{\boldsymbol{e}}_i, Isomap(F_{\boldsymbol{\theta}_1}(\boldsymbol{x}_i)))|_{i=1}^{N}). \tag{6}$$

Unfortunately, it is difficult to integrate the Isomap into the back propagation process because it is both time and space consuming. The time complexity of Isomap is $O(N^2 log_N)$ and the space demand is $O(N^2)$, where N is the number of samples. To solve this issue, we propose the Isometric Propagator $IP_{\boldsymbol{\theta}_3}(\cdot)$ to parameterize the Isomap algorithm. Isometric Propagator is a three layer MLP trained to simulate the Isomap function at the beginning of Sphere-Oriented Training. We freeze the parameter of the pretrained CNN $F_{\boldsymbol{\theta}_1}$ and train the Isometric Propagator as follow:

$$\arg\min_{\boldsymbol{\theta}_3}(\mathcal{L}_1(Isomap(\boldsymbol{f}_i), IP_{\boldsymbol{\theta}_3}(\boldsymbol{f}_i))|_{i=1}^{N}). \tag{7}$$

After the training of the Isometric Propagator, we freeze its parameters and replace the Isomap with it to train the feature extractor CNN. The actual Sphere-Oriented Training is formalized as:

$$\arg\min_{\boldsymbol{\theta}_1}(\mathcal{L}_1(\hat{\boldsymbol{e}}_i, IP_{\boldsymbol{\theta}_3}(F_{\boldsymbol{\theta}_1}(\boldsymbol{x}_i)))|_{i=1}^{N}). \tag{8}$$

Note that the Isometric Propagator is only used during the source domain training. At test time, we predict gaze

by the proposed GPM with Isomap for better generalization ability. Parameters of the GPM are determined before the Sphere-Oriented Training and remain fixed.

Owing to the advantage that the GPM suffers less from the overfitting problem than the FC layer, the purpose of the SOT is to utilize this advantage to optimize the gaze estimation model for better generalization performance by incorporating the GPM into the training process.

4.3. Implementation Details

We employ the AGG by PyTorch. For the training of the pretrain model, IP and Sphere-Oriented Training, we use the Adam optimizer with a learning rate of 10^{-4}. The model is pretrained for 10 epochs. We choose the last epoch as the baseline model. The Sphere-Oriented Training is also 10 epochs, while the IP is trained for 100 epochs on 2000 randomly selected samples. Batch sizes are set to 512. For Isomap, we use the implementation of Scikit-learn and the number of neighbor is set to 300. Pixel values are normalized to $[0, 1]$, and no data augmentation is employed.

5. Experiments

5.1. Data Preparation

We conduct experiments on four commonly used gaze estimation datasets: ETH-XGaze ($\mathcal{D}_E$) [39], Gaze360 ($\mathcal{D}_G$) [12], MPIIFaceGaze ($\mathcal{D}_M$) [38] and EyeDiap ($\mathcal{D}_D$) [9]. We normalize the data following the techniques in [38]. **ETH-XGaze:** $756k$ images captured by high resolution cameras in laboratory environment with large gaze range. We divide the last 5 subjects as test set. **Gaze360:** $101k$ images captured by a 360° camera on streets with large gaze range. We only use images with frontal faces in our experiments. **MPIIFaceGaze:** $45k$ images (standard test set) captured by web camera during daily usage of laptop computers. The gaze range of $\mathcal{D}_M$ is less than half the range of $\mathcal{D}_E$ and $\mathcal{D}_G$. Thus, we only use $\mathcal{D}_M$ as target domain. **EyeDiap:** $16k$ images captured under laboratory environment with screen and floating targets. As the number of images is significantly less than other datasets, we only use $\mathcal{D}_D$ as target domain.

In addition, the cross-domain error between $\mathcal{D}_E$ and $\mathcal{D}_G$ is extremely large (around 20°). Thus, we exclude the $\mathcal{D}_E \rightarrow \mathcal{D}_G$ and $\mathcal{D}_G \rightarrow \mathcal{D}_E$ settings in our experiments, which is also excluded in previous studies [2, 8, 20, 33].

5.2. Quantitative Evaluation

5.2.1 Evaluation of the GPM

We first evaluate the proposed GPM by replacing the last FC layer with the GPM without changing other parameters of the baseline model. The mean and the standard deviation (std) of the estimation error from the last 5 epochs are shown in Tab. 1. The proposed GPM achieves better performance in 7 out of 8 cross-domain experiments. In addition, the GPM also performs more stably across different epochs. These results demonstrate the advantage of the proposed GPM over the traditional FC layer. The within-dataset estimation errors of the GPM are slightly higher. It is reasonable since GPM is designed for generalizing to unseen domains. The higher within-dataset accuracy of the FC layer is highly likely achieved by overfitting since it performs worse in cross-domain tests.

Table 1. Results of simply replacing the last FC layer with the proposed GPM in inference. Results are the mean and std for the final 5 epochs. Note that the modest reduction in within-dataset accuracy is reasonable, since GPM is designed for generalization.

	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	within $\mathcal{D}_E$
ResNet-18	8.66±0.53	7.76±0.29	**5.37**±0.24
ResNet-18 + GPM	**7.87**±0.23	**7.72**±0.33	5.74±0.08
ResNet-50	6.92±0.86	8.61±0.88	**5.27**±0.56
ResNet-50 + GPM	**6.56**±0.41	**8.10**±0.57	5.29±0.07
	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$	within $\mathcal{D}_G$
ResNet-18	8.59±0.57	**10.87**±1.52	**12.59**±0.14
ResNet-18 + GPM	**8.57**±0.41	10.94±0.85	12.64±0.10
ResNet-50	8.48±1.01	10.76±0.78	**11.97**±0.30
ResNet-50 + GPM	**8.14**±0.47	**9.77**±1.00	12.07±0.18

5.2.2 Evaluation of the AGG Framework

In this section, we evaluate the effectiveness of the AGG framework, which optimizes the gaze estimation model to further improve generalization ability. We conduct experiments in 4 cross domain settings with 3 baseline models, as shown in Tab. 2. The performances of baseline models is quite different, due to their architectures and the different characteristics of each domain. Nevertheless, the proposed AGG framework achieves stable improvements in all 12 cross-domain settings, proves that the AGG framework is robust to different baseline models and source domains. The AGG framework achieves improvements as large as 35.79% without target domain data. We also report the within dataset performance after generalization for reference. As expected, the within dataset performance decreases mildly since the model is optimized for domain generalization. Above results demonstrate the effectiveness of the proposed Sphere-Oriented Training, which improves the generalization performance of varies baseline models significantly.

5.2.3 Comparison with SOTA Methods

In Tab. 3, we compare the AGG framework with SOTA gaze estimation methods [6, 12, 38] and gaze general-

Table 2. Performance of the proposed AGG framework. Results are gaze estimation error in degrees. The proposed AGG achieves stable improvements up to 35.79% in all 12 cross-domain settings without using any target domain data. The symbol * indicates within-dataset experiments for reference. Note that the modest reduction in within-dataset accuracy is to be expected for domain generalization methods.

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$	within $\mathcal{D}_E$*	within $\mathcal{D}_G$*
ResNet-18	8.64	7.83	8.68	12.35	5.08	12.73
ResNet-18+AGG	7.10 ▼ 17.82%	7.07 ▼ 9.71%	7.87 ▼ 9.33%	7.93 ▼ 35.79%	5.56 ▲ 9.45%	13.03 ▲ 2.36%
ResNet-50	6.04	7.47	10.14	11.76	5.35	12.37
ResNet-50+AGG	5.91 ▼ 2.15%	6.75 ▼ 9.64%	9.2 ▼ 9.27%	11.36 ▼ 3.40%	6.29 ▲ 17.57%	15.63 ▲ 26.35%
VGG16	9.5	19.14	14.61	19.94	5.12	12.15
VGG16+AGG	9.13 ▼ 3.89%	17.2 ▼ 10.14%	11.3 ▼ 22.66%	13.97 ▼ 29.94%	5.78 ▲ 12.89%	13.13 ▲ 8.07%

Table 3. Cross domain gaze estimation error in degrees. $*$ indicates methods with ResNet-50 backbone. Overall, the proposed AGG achieves better generalization ability than SOTA gaze estimation methods.

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
Full-Face[38]	12.35	30.15	11.13	14.42
ADL[12]	7.23	8.02	11.36	11.86
CA-Net[6]	-	-	27.13	31.41
LatentGaze[18]	7.98	9.81	-	-
PureGaze[8]	7.08*	7.48*	9.28	9.32
ResNet18+AGG	7.10	7.07	**7.87**	**7.93**
ResNet50+AGG	**5.91***	**6.75***	9.20*	11.36*

ization methods [8, 18]. Results demonstrate that the AGG outperforms other SOTA methods. The AGG with ResNet-18 baseline achieves the best overall performances, it outperforms SOTA methods in $\mathcal{D}_E \rightarrow \mathcal{D}_D$, $\mathcal{D}_G \rightarrow \mathcal{D}_M$ and $\mathcal{D}_G \rightarrow \mathcal{D}_D$ settings, while achieving performance comparable to the PureGaze in the $\mathcal{D}_E \rightarrow \mathcal{D}_M$ setting. The AGG with ResNet-50 baseline also surpasses SOTA methods in 3 out of 4 cross-domain settings. It performs exceptionally well when trained in the $\mathcal{D}_E$ domain. Overall, above experiments prove that the AGG achieves better generalization ability than SOTA gaze estimation methods.

5.3. Verification of the AGG

5.3.1 Verification of the Core Idea

The proposed Analytical Gaze Generalization framework is designed based on the observation that the geodesic distance between image features is proportional to the angular gaze differences between input samples. To explore whether this observation holds true in different domains, we verify it in $\mathcal{D}_E$, $\mathcal{D}_G$, $\mathcal{D}_M$ and $\mathcal{D}_D$ respectively. We train a baseline ResNet-18 model according to Eq. (1) in each domain respectively to extract the image feature, and visualize the L2 and Geodesic distance with respect to the angular gaze differences. As shown in Fig. 6, the linear relationship holds true for all sample pairs in $\mathcal{D}_E$, thanks to the high image quality and controlled laboratory environment. For $\mathcal{D}_G$, the pattern is evident at the beginning but becomes random

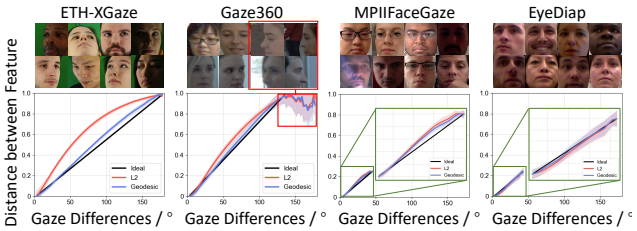

Figure 6. The L2 and Geodesic distances between image features with respect to the angular differences between samples.

when gaze differences surpass 140°. We randomly visualize 4 samples from the random section at the top of the figure. Since the original $\mathcal{D}_G$ dataset includes subjects facing away from the camera, the quality of samples appears to deteriorate when the head pose approaches $\pm 90°$. For $\mathcal{D}_M$ and $\mathcal{D}_D$, the geodesic distance is also exhibits a more direct proportionality to gaze differences. However, the disparity between geodesic distance and L2 distance is less obvious compared to what was observed in the $\mathcal{D}_E$ and $\mathcal{D}_G$. It is reasonable since the gaze ranges in $\mathcal{D}_M$ and $\mathcal{D}_D$ are significantly smaller. The geodesic distance converges toward the L2 distance when the features are in close proximity.

We further visualize the Principle Gaze Feature from these four datasets in Fig. 7 for a more intuitive understanding. The Principle Gaze Feature from all 4 datasets consistently demonstrate identical distribution pattern with the gaze label. Above results confirm that the the proportional relationship between the geodesic distance and the gaze differences remains consistent across different domains, even though the image quality, gaze range, and head pose range exhibit significant variations among these domains. Hence, the proportional relationship can be employed for domain generalization, given that it is domain-independent.

5.3.2 Verification of the Sphere-Oriented Training

In this section, we assess the efficacy of the Sphere-Oriented Training, *i.e.* whether Sphere-Oriented Training optimizes the model to extract features that better conform to the pro-

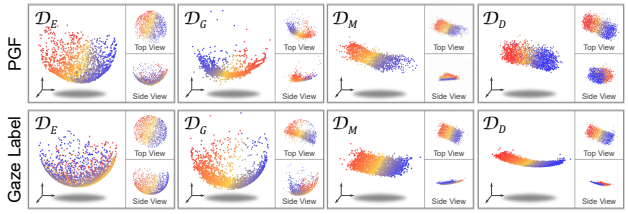

Figure 7. Visualization of the Principle Gaze Feature (PGF) and gaze label from $\mathcal{D}_E$, $\mathcal{D}_G$, $\mathcal{D}_M$ and $\mathcal{D}_D$. PGF from all 4 datasets share the same distribution pattern with gaze label.

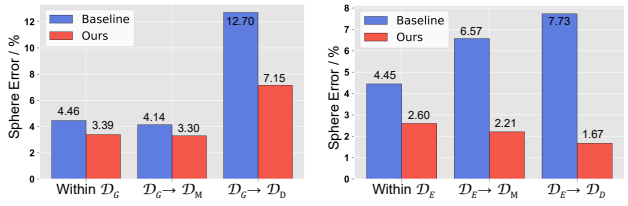

Figure 8. The Sphere Error of the PGF before (ResNet-18) and after Sphere-Oriented Training (Ours). Smaller sphere error indicates that the geodesic distance between extracted features are more proportional to the gaze differences.

portional relationship. To do so, we measure the Sphere Error, defined as the ratio of distance between e_i and the sphere surface to the radius of the sphere. The quantitative results presented in Fig. 8 demonstrate that the Sphere Error after the Sphere-Oriented Training reduces in both within-domain and cross-domain settings. Fig. 9 provides a more intuitive view. The Sphere Error is significantly reduced in the central region after the Sphere-Oriented Training and the distribution of the PGF becomes more spherical. Above results validates the effectiveness of the propsoed Sphere-Oriented Training.

6. Limitations and Discussions

Q1: Does the observation in Fig. 2 apply to features extracted by different gaze estimation models? We have proven that the observation holds true for different model architectures in Tabs. 1 and 2 and different datasets Fig. 7. Here we further investigate the influence of two loss functions: $\mathcal{L}_1, \mathcal{L}_2$ loss, and two gaze representations: 3D unit vector (x, y, z), 2D Euler angle $(yaw, pitch)$. In Fig. 10, we train ResNet-18 models under above conditions and project the extracted features to 3D space using geodesic distance. Results show that the AGG is robust to different loss functions. When gaze is represented by 2D Euler angles, the projected features no longer distributes across the sphere surface, they approximately distribute on the surface of a 2D plane, similar to the gaze label. The Sphere Alignment algorithm needs to be altered to adapt different gaze representations, which we have left for future work.

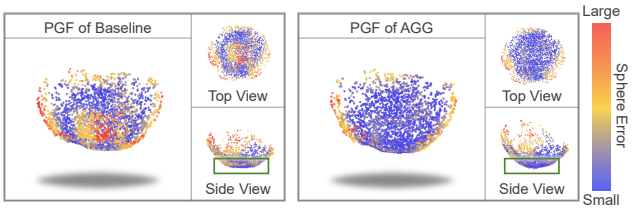

Figure 9. Visualization of the Sphere Error before and after Sphere-Oriented Training in $\mathcal{D}_E$. In the side view, the bottom area of the PGF of AGG is more spherical.

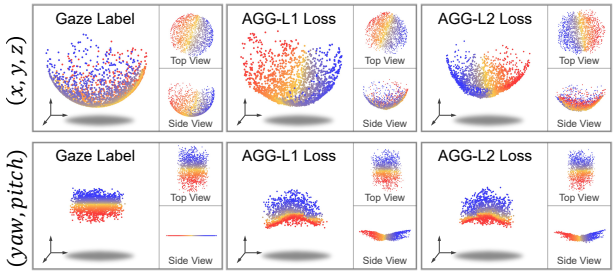

Figure 10. Projecting features extracted by ResNet-18 trained with different strategy in $\mathcal{D}_E$ into 3D space by geodesic distance.

Q2: Does the Isometric Propagator (IP) suffer from the same overfitting issue as the last FC layer since it is implemented by MLP? Although we completely replace the last FC layer with the GPM in the test time, the MLP based IP is still used in source domain training. The use of IP is an unavoidable compromise, because in the current deep-learning community, there has not been a perfect solution for integrating Isomap into the back-propagation process. But the IP differs from the original regression FC layer, since IP is only used in the training time, as a tool for back-propagation with fixed parameters. Still, the implementation of IP is a limitation to our method. The performance of the AGG might be further improved if there is a better way to integrate Isomap into the training process.

7. Conclusion

In this paper we propose the Analytical Gaze Generalization framework for generalizing gaze estimation models to unseen domains. Based on the observation that the geodesic distance between extracted image features is proportional to the angular gaze differences, we propose the Geodesic Projection Module that estimates gaze from the image feature analytically and incorporate it into the source domain training by the proposed Sphere-Oriented Training. Extensive experiments show that the GPM achieves better generalization ability than the conventional FC layer, and the AGG improves the cross-domain accuracy significantly, outperforming SOTA methods. The concept of the AGG may inspire method designs in other physical regression tasks, *e.g.* pose estimation.

References

[1] Yiwei Bao, Yihua Cheng, Yunfei Liu, and Feng Lu. Adaptive feature fusion network for gaze tracking in mobile tablets. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9936–9943. IEEE, 2021. 2

[2] Yiwei Bao, Yunfei Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with rotation consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4207–4216, 2022. 2, 6

[3] Alisa Burova, John Mäkelä, Jaakko Hakulinen, Tuuli Keskinen, Hanna Heinonen, Sanni Siltanen, and Markku Turunen. Utilizing vr and gaze tracking to develop ar solutions for industrial maintenance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020. 1

[4] Nora Castner, Thomas C Kuebler, Katharina Scheiter, Juliane Richter, Thérése Eder, Fabian Hüttig, Constanze Keutel, and Enkelejda Kasneci. Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–10, 2020. 1

[5] Zhaokang Chen and Bertram E Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer, 2018. 2

[6] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10623–10630, 2020. 2, 6, 7

[7] Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272, 2020. 2

[8] Yihua Cheng, Yiwei Bao, and Feng Lu. Puregaze: Purifying gaze feature for generalizable gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 436–443, 2022. 1, 2, 3, 6, 7

[9] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. 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*, pages 255–258, 2014. 2, 6

[10] Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on biomedical engineering*, 53(6):1124–1133, 2006. 2

[11] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500, 2009. 2

[12] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921, 2019. 1, 2, 6, 7

[13] Jess Kerr-Gaffney, Amy Harrison, and Kate Tchanturia. Eye-tracking research in eating disorders: A systematic review. *International Journal of Eating Disorders*, 52(1):3–27, 2019. 1

[14] Andrew J King, Gregory F Cooper, Gilles Clermont, Harry Hochheiser, Milos Hauskrecht, Dean F Sittig, and Shyam Visweswaran. Leveraging eye tracking to prioritize relevant medical record data: comparative machine learning study. *Journal of medical Internet research*, 22(4):e15876, 2020. 1

[15] Robert Konrad, Anastasios Angelopoulos, and Gordon Wetzstein. Gaze-contingent ocular parallax rendering for virtual reality. *ACM Transactions on Graphics (TOG)*, 39(2):1–12, 2020. 1

[16] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016. 2

[17] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A Lee, and Mark Billinghurst. Pinpointing: Precise head-and eye-based target selection for augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018. 1

[18] Isack Lee, Jun-Seok Yun, Hee Hyeon Kim, Youngju Na, and Seok Bong Yoo. Latentgaze: Cross-domain gaze estimation through gaze-aware analytic latent code manipulation. In *Proceedings of the Asian Conference on Computer Vision*, pages 3379–3395, 2022. 7

[19] Jiahui Liu, Jiannan Chi, and Shuo Fan. A method for accurate 3d gaze estimation with a single camera and two collinear light sources. *IEEE Transactions on Instrumentation and Measurement*, 2022. 2

[20] Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3835–3844, 2021. 1, 2, 6

[21] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Adaptive linear regression for appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 36(10):2033–2046, 2014. 2

[22] Feng Lu, Xiaowu Chen, and Yoichi Sato. Appearance-based gaze estimation via uncalibrated gaze pattern recovery. *IEEE Transactions on Image Processing*, 26(4):1543–1553, 2017. 2

[23] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9368–9377, 2019. 2

[24] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290 (5500):2323–2326, 2000. 4

[25] Timo Schneider, Boris Schauerte, and Rainer Stiefelhagen. Manifold alignment for person independent appearance-based gaze estimation. In *2014 22nd international conference on pattern recognition*, pages 1167–1172. IEEE, 2014. 2

[26] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4):1633–1642, 2018. 1

[27] Sophie Stellmach, Sebastian Stober, Andreas Nürnberger, and Raimund Dachselt. Designing gaze-supported multimodal interactions for the exploration of large image collections. In *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications*, New York, NY, USA, 2011. Association for Computing Machinery. 1

[28] Li Sun, Zicheng Liu, and Ming-Ting Sun. Real time gaze estimation with a consumer depth camera. *Information Sciences*, 320:346–360, 2015. 2

[29] Kentaro Takemura and Kenta Yamagishi. A hybrid eye-tracking method using a multispectral camera. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1529–1534. IEEE, 2017. 2

[30] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000. 4

[31] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 4

[32] Kang Wang, Rui Zhao, Hui Su, and Qiang Ji. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11907–11916, 2019. 1

[33] Yaoming Wang, Yangzhou Jiang, Jin Li, Bingbing Ni, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Teng Li. Contrastive regression for domain adaptation on gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19376–19385, 2022. 1, 2, 6

[34] Zhimin Wang, Huangyue Yu, Haofei Wang, Zongji Wang, and Feng Lu. Comparing single-modal and multimodal interaction in an augmented reality system. In *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct, ISMAR 2020 Adjunct, Recife, Brazil, November 9-13, 2020*, pages 165–166. IEEE, 2020. 1

[35] Xuehan Xiong, Zicheng Liu, Qin Cai, and Zhengyou Zhang. Eye gaze tracking using an rgbd camera: A comparison with a rgb solution. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 1113–1121, 2014. 2

[36] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7314–7324, 2020. 2

[37] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015. 2

[38] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 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*, pages 51–60, 2017. 2, 6, 7

[39] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, pages 365–381. Springer, 2020. 2, 3, 6