This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

See Through the Noise: Improving Domain Generalization in Gaze Estimation

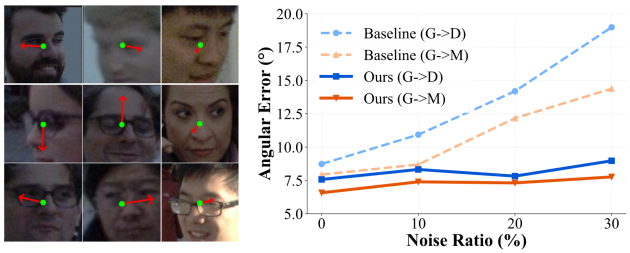
Yanming Peng, Shijing Wang, Yaping Huang*, Yi Tian
Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing Jiaotong University
pengym@bjtu.edu.cn, shijingwang@bjtu.edu.cn, yphuang@bjtu.edu.cn, tianyi@bjtu.edu.cn

Abstract

Generalizable gaze estimation methods have garnered increasing attention due to their critical importance in real-world applications and have achieved significant progress. However, they often overlook the effect of label noise, arising from the inherent difficulty of acquiring precise gaze annotations, on model generalization performance. In this paper, we are the first to comprehensively investigate the negative effects of label noise on generalization in gaze estimation. Further, we propose a novel solution, called See-Through-Noise (SeeTN) framework, which improves generalization from a novel perspective of mitigating label noise. Specifically, we propose to construct a semantic embedding space via a prototype-based transformation to preserve a consistent topological structure between gaze features and continuous labels. We then measure feature-label affinity consistency to distinguish noisy from clean samples, and introduce a novel affinity regularization in the semantic manifold to transfer gaze-related information from clean to noisy samples. Our proposed SeeTN promotes semantic structure alignment and enforces domain-invariant gaze relationships, thereby enhancing robustness against label noise. Extensive experiments demonstrate that our SeeTN effectively mitigates the adverse impact of source-domain noise, leading to superior cross-domain generalization without compromising the source-domain accuracy, and highlight the importance of explicitly handling noise in generalized gaze estimation.

1. Introduction

Gaze estimation techniques have been extensively applied in various fields, including human-computer interaction [16], virtual reality [5, 15], and medical analysis [7]. Recently, powered by deep learning technologies, appearance-based gaze estimation [4, 21, 31] has achieved significant performance improvement. However, existing models often suffer from substantial performance degradation when deployed to unseen domains. To enhance generalization in real-world scenarios, recent works have focused on learning domain-invariant representations to handle environmental variations via adversarial learning [9, 23], contrastive learning [25], consistency constraints [3, 27], *etc.*



(a) Noisy Samples (b) Preliminary Experiments

Figure 1. (a) Noisy samples from the Gaze360 dataset, where red arrows indicate the provided ground-truth labels that clearly deviate from the true gaze directions. (b) Cross-domain results on the synthetic noisy Gaze360 dataset under cross-domain settings, Eyediap (G→D) and MPIIGaze (G→M). The baseline (ResNet-18) performance degrades significantly with increasing noise, while our method consistently maintains robust generalization.

Although these methods have achieved remarkable progress, they often overlook the widely existing noisy labels in the gaze estimation task. In practice, gaze estimation datasets are notoriously difficult to collect due to several factors, including unstable human attention and environmental variations (*e.g.*, lighting, head pose, and occlusion). These challenges lead to inevitable annotation noise in large-scale gaze datasets. As shown in Fig. 1(a), a considerable amount of noise exists in the commonly used Gaze360 dataset [14]. Intuitively, overfitting to noisy labels forces the model to inadvertently learn gaze-irrelevant features, thereby degrading its generalization performance. To comprehensively explore the impact of noisy labels in cross-domain scenarios, we conduct preliminary experiments on the synthetic noisy dataset by injecting label noise into the Gaze360 dataset. In detail, we randomly select

*Corresponding author.

10%, 20%, and 30% of the samples and add Gaussian noise with a standard deviation of 60° to their labels. As shown in Fig. 1(b), we can observe that label noise adversely affects the model's generalization performance, *e.g.*, the performance degrades from 7.94° to 12.17° under 20% and further to 14.38° under 30% label noise on MPIIGaze. Moreover, this is also validated by the results of real-world experiments (see Sec. 4.3 for details).

Unfortunately, most existing noisy label learning methods mainly focus on mitigating within-domain performance degradation in classification tasks, but the challenge we aim to address is a regression task with cross-domain degradation. This raises two key issues that remain unexplored in related studies and thus need to be addressed. First, compared with classification tasks, dealing with label noise in regression is substantially more challenging due to the continuous nature of the target space. In regression, label noise is not categorical misclassification but continuous deviation, making it difficult to define clear boundaries between clean and noisy samples. Moreover, the feature space often lacks distinct clusters, and common classification-based heuristics such as small-loss selection or transition-matrix estimation become unreliable. Second, due to the inherent difficulty of cross-domain scenarios, simply correcting noisy labels does not necessarily improve generalization performance, because such correction primarily refines the label quality within the source domain but fails to bridge the distribution gap between the source and target domains.

In this paper, we address the domain generalization in gaze estimation from a novel perspective of mitigating noise. We propose a **See**-**T**hrough-**N**oise framework, called SeeTN, which models the semantic relatedness between continuous gaze labels and feature representations to effectively identify noisy samples and mitigate domain distribution discrepancies. This is particularly crucial for gaze regression tasks, where labels are inherently continuous. Specifically, we propose a manifold representation of the semantic space using a prototype-based transformation, which can preserve and constrain the relative relationships among semantic features according to the label affinity. Through the proposed semantic manifold, we can measure the discrepancy of features and labels, which greatly facilitates the distinction between noisy and clean samples. Built upon the separation, we further propose a novel affinity regularization in semantic manifold space, where different losses are respectively applied to clean and noisy samples, thereby resulting in more robust regularization of gaze-relevant features. Comprehensive experimental results show that SeeTN effectively mitigates the negative impacts of noisy labels and improves the model's domain generalization ability. In conclusion, our contributions can be summarized in three aspects:

- We present the first work to investigate and address the domain generalization of gaze estimation from a novel perspective of label noise, arising inevitably during the process of gaze dataset acquisition.
- We propose a novel See-Through-Noise framework, named SeeTN, which bridges continuous gaze angle labels and gaze features in a semantic manifold constructed via a prototype-based transformation. By leveraging the manifold space, we measure the discrepancy between feature and label affinities to identify noisy samples, and further design a novel affinity regularization strategy for robust learning of clean and noisy samples.
- We conduct comprehensive experiments on various cross-domain settings. The results on real-world datasets demonstrate that SeeTN effectively mitigates the negative impacts of noisy labels and learns more consistent semantic information.

2. Related Works

2.1. Cross-domain Gaze Estimation

The challenge of enabling gaze estimation models trained in controlled laboratory settings to perform reliably in complex real-world scenarios has led to growing interest in cross-domain gaze estimation. Generally, cross-domain gaze estimation can be categorized into two settings: unsupervised domain adaptation and domain generalization.

Unsupervised domain adaptation assumes that only a limited amount of unlabeled target domain data is accessible. CRGA [25] utilizes contrastive regression learning to acquire better feature representations. RUDA [3] adapts to the target domain through rotation consistency of gaze. UnReGA [6] proposes source-free unsupervised domain adaptation in gaze estimation. Domain generalization, which cannot access target domain data, is more versatile and challenging. PureGaze [9] purifies gaze features to improve generalization performance. AGG [2] mitigates the overfitting during the training stage by replacing the last FC layer. GazeCF [29] leverages synthetic data to purify gaze-related features. CGaG [27] employs collaborative contrastive learning to learn the gaze-related features.

Although both adaptation and generalization methods have achieved remarkable performance, they have largely overlooked the potential noise in gaze labels of the source domain, which serve as the supervision signals and provide the domain-invariant semantic information in cross-domain gaze estimation. In contrast, we propose a novel perspective of handling noisy labels to enhance the domain generalization performance.

2.2. Noisy Label Learning

The mainstream methods of noisy label learning (LNL) that mainly focus on classification tasks can be roughly categorized into loss correction and sample selection. For

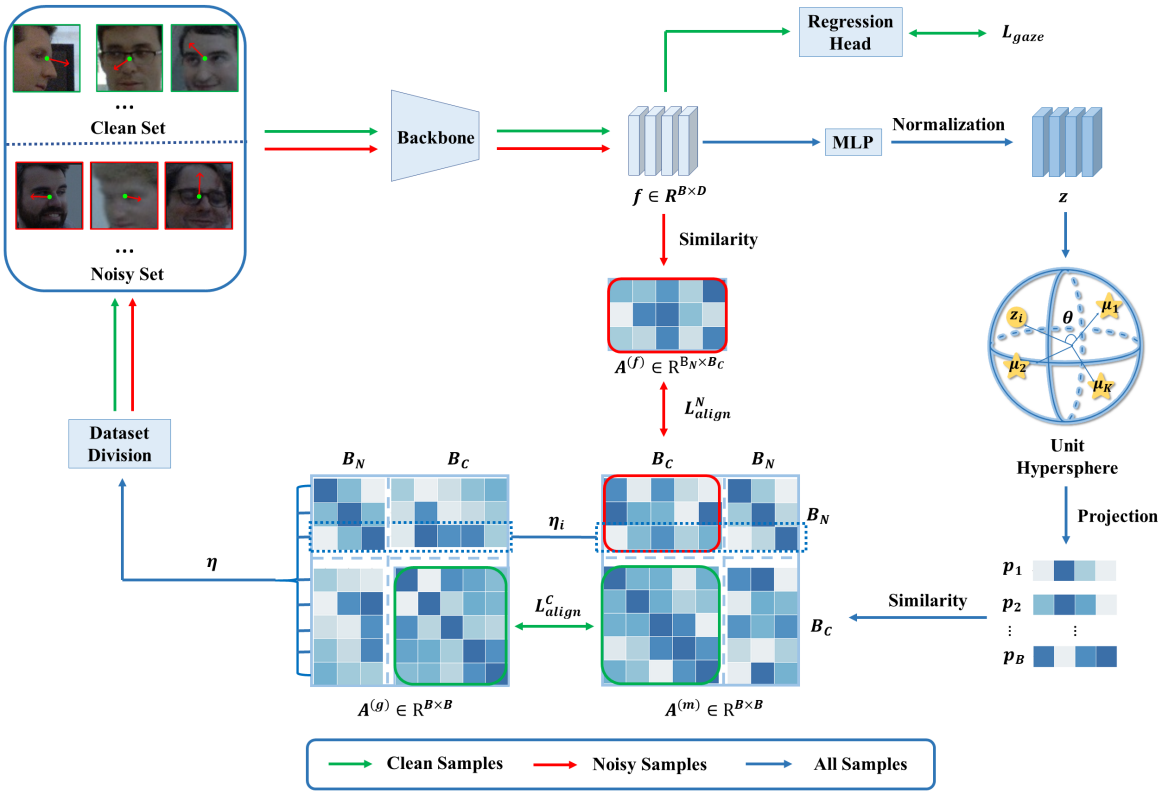


Figure 2. Overview of the proposed SeeTN, which consists of three main steps. First, SeeTN constructs the semantic manifold to strengthen the associations between gaze features and their corresponding continuous gaze labels. Second, according to the semantic discrepancies between features and their labels in the manifold, SeeTN divides the dataset into a noisy set and a clean set. Finally, to better utilize noisy samples and clean ones, SeeTN introduces a noise-robust regularization mechanism that constrains the noisy and clean sets separately. In the figure, red and green denote noisy and clean samples, respectively.

loss correction, a common strategy is to estimate the noise transition matrix and design regularized loss functions [1, 28]. Recently, benefiting from the development of semi-supervised learning techniques, sample selection methods have become more popular. Many approaches [18, 22, 26] divide datasets into clean and noisy subsets via sample selection strategies and leverage semi-supervised learning to generate pseudo-labels and improve performance. Small-loss criterion [12, 30, 35] is commonly used in sample selection due to the memorization effect that DNNs are more likely to fit clean samples in early training.

However, LNL is predominantly designed for classification tasks, whereas gaze estimation is a regression task. We have empirically demonstrated that simply transferring LNL techniques to gaze estimation is not enough. To the best of our knowledge, in gaze estimation, only SUGE [24] first utilizes the co-training strategy, estimates the uncertainty for suppressing noisy samples, and generates pseudo-labels via neighbor labeling. Nevertheless, SUGE addresses only the within-domain setting, overlooking the more practical cross-domain scenario. In cross-domain settings, addressing both noise and generalization simultaneously for gaze estimation poses a significant challenge.

3. Method

3.1. Problem Definition

We denote the source domain data as $\mathcal{D}_S = \{(x_i, y_i)\}_{i=1}^{N_S}$, where (x_i, y_i) corresponds to an image and its 3D gaze label. Let F and G denote the feature extractor and regressor of DNNs respectively, and the prediction of the image x is $G(F(x;\theta_f);\theta_g)$, where θ_f and θ_g denote the parameters of the model. To simplify the formulations, we use $f(x) = F(x;\theta_f)$ and $g(x) = G(F(x;\theta_f);\theta_g)$ as alternatives. For domain generalization, typically, the objective function in the source domain can be formulated as:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)}[\mathcal{L}_1(g(x), y)], \tag{1}$$

and the model with θ^* is directly used to test the target domain. Since x and y contain low-quality images and labels, the model could overfit the noise, leading to poor domain generalization performance.

3.2. Overview

Fig. 2 shows the overall framework of the proposed SeeTN. SeeTN aims to enhance the semantic association between gaze features and continuous gaze angles, which facilitates preserving the geometric structure of the semantic space in the gaze regression task. Meanwhile, the semantic discrepancies between labels and features observed during the association process can assist the model in identifying noisy samples. To achieve the goal, SeeTN first constructs a semantic manifold from feature-derived prototypes, in which the affinity constraints between features in the manifold can be formulated based on the similarity between their corresponding labels. By quantifying the discrepancy of features and labels, an indicator is designed for identifying noisy labels. For noisy samples with unreliable labels, SeeTN regularizes the raw features by leveraging the affinity matrix in the manifold, which seeks to transfer the gaze-related information supervised by the clean samples to noisy samples. Through tailored regularization for clean and noisy samples, our model strengthens the learning of gaze-relevant representations, leading to the improvement of domain generalization. In the following sections, we present the proposed method in detail.

3.3. Semantic Manifold Construction

In regression tasks, the commonly used mean absolute error (MAE) loss focuses on the numerical accuracy of predictions, while overlooking the semantic structures in both the feature and label spaces. This implies that samples with similar labels may still have feature representations scattered across the feature space, which is detrimental to the semantic correspondence between features and labels. For better identifying noisy samples and improving model generalization, it is essential to strengthen the semantic alignment between features and labels. Thus motivated, we propose to explicitly construct a semantic manifold $\mathcal{M}$ to model the semantic structures of features.

Formally, to emphasize gaze semantics and suppress irrelevant information, we first transform the feature f into another space through MLP and project it onto the unit hypersphere as $z = \text{Norm}(\text{MLP}(f))$. Then, K prototypes, denoted as $\mu_k, k = 1, 2, \ldots, K$, are randomly selected in the unit hypersphere to serve as the basis vectors of $\mathcal{M}$. To ensure that all prototypes μ_k contain more representative semantic information, they are progressively moved toward the feature center via an Exponential Moving Average (EMA) update:

$$r = \text{Softmax}(z \cdot \mu^T / \tau), \tag{2}$$

$$\mu_k \leftarrow \alpha\, \mu_k + (1-\alpha)\frac{\sum_i r_{i,k} z_i}{\sum_i r_{i,k}}, \tag{3}$$

where $r \in \mathbb{R}^{B\times K}$ denotes the assignment probabilities of z with respect to the prototypes μ, B is the batch size, and α represents the EMA momentum and is set to 0.95. In addition, relying on the semantic prototypes μ, the representation p of z on the manifold $\mathcal{M}$ can be expressed:

$$p = z \cdot \mu^{\mathrm{T}}. \tag{4}$$

Motivated by the continuous rather than categorical nature of gaze angles, we constrain the learned semantic manifold to preserve a topological structure consistent with that of the label space. Therefore, we first measure the affinity matrix in the manifold $\mathcal{M}$ denoted as $A^{(m)}$, which represents the relative relationships between samples defined by cosine similarity:

$$A^{(m)}_{i,j} = \frac{p_i \cdot p_j}{\|p_i\|\|p_j\|}. \tag{5}$$

The gaze label affinity matrix $A^{(g)}$ can also be calculated by

$$A^{(g)}_{i,j} = \frac{y_i \cdot y_j}{\|y_i\|\|y_j\|}. \tag{6}$$

Then we can align the gaze label affinity matrix $A^{(g)}$ and the manifold feature affinity matrix $A^{(m)}$ for establishing a bridge between them on the $\mathcal{M}$. By minimizing the MAE $|A^{(m)} - A^{(g)}|$ between manifold feature and label affinity matrix, the manifold $\mathcal{M}$ is encouraged to emphasize the semantic relationships between samples, while reinforcing the similarity-based associations between features and labels, thus providing a solid basis for the subsequent noise filtering and robust learning process.

3.4. Noise-Robust Regularization

In gaze estimation, noisy samples are inevitable, as collecting and annotating large-scale datasets is inherently challenging. To alleviate the negative impact of noisy samples and improve the generalization abilities, we propose to distinguish noisy samples from clean ones. Although cross entropy loss-based clean sample selection is not feasible for regression tasks, by exploiting the semantic manifold, we can quantify the affinity mismatch between features and labels using the cross-entropy function as a criterion. Moreover, we further design a novel affinity regularization to improve the model generalization.

Specifically, intuitively, the normalized row vector $\hat{y}^{(m)}_i = \text{Softmax}(A^{(m)}_{i,:})$ in the semantic manifold $\mathcal{M}$ can be interpreted as a distribution of sample x_i based on its relations to the remaining samples. Similarly, the row vector $\hat{y}^{(g)}_i = \text{Softmax}(A^{(g)}_{i,:})$ can also be regarded as the distribution of sample x_i over all other samples in the gaze label space. Therefore, for each sample x_i, we can introduce an indicator η_i to describe the discrepancy between its feature and label as follows:

$$\eta_i = -\sum_{j=1}^{B} \hat{y}^{(g)}_{i,j} \log \hat{y}^{(m)}_{i,j}. \tag{7}$$

By integrating features, labels and the relative relationships among samples, the indicator η_i possesses the capability to identify noise. Noisy samples tend to exhibit inconsistencies between the pairwise relationships in the label space and those in the manifold feature space, therefore, samples with larger η can be regarded as noisy samples. To divide samples into clean and noisy ones, we adopt a simple strategy by sorting all samples in descending order according to η, and selecting the top $t\%$ as the noisy subset $\mathcal{D}_{\mathcal{S}}^{\mathcal{N}}$, while the remaining samples are regarded as the clean subset $\mathcal{D}_{\mathcal{S}}^{\mathcal{C}}$.

Admittedly, this simple partitioning strategy may introduce risks. $\mathcal{D}_{\mathcal{S}}^{\mathcal{N}}$ might contain some hard samples that have not yet been well learned by the model, and $\mathcal{D}_{\mathcal{S}}^{\mathcal{C}}$ may still include a few noisy samples with small deviations. Nevertheless, this approach places most noisy samples and low-quality images in the noisy subset, which is beneficial for improving model robustness. Meanwhile, before each training epoch, we repartition the dataset according to the updated η to mitigate confirmation bias.

Based on the above partition, we further design a noise-robust regularization item that applies distinct constraints to clean and noisy samples separately. For $\mathcal{D}_{\mathcal{S}}^{\mathcal{C}}$, as the labels are more reliable, we directly use them to supervise the gaze estimation loss and constrain the affinity matrix $A^{(m)}$ within the manifold space $\mathcal{M}$:

$$\mathcal{L}_{\text{gaze}} = \frac{1}{B_C} \sum_{x_i \in \mathcal{D}_{\mathcal{S}}^{\mathcal{C}}} |g(x_i) - y_i|, \tag{8}$$

$$\mathcal{L}_{\text{align}}^{\mathcal{C}} = \frac{1}{B_C(B_C-1)} \sum_{x_i \in \mathcal{D}_{\mathcal{S}}^{\mathcal{C}}} \sum_{x_j \in \mathcal{D}_{\mathcal{S}}^{\mathcal{C}}, i \neq j} |A_{i,j}^{(g)} - A_{i,j}^{(m)}|, \tag{9}$$

where B_C is the batch size of $\mathcal{D}_{\mathcal{S}}^{\mathcal{C}}$. For $\mathcal{D}_{\mathcal{S}}^{\mathcal{N}}$, a simple discard would lead to the loss of valuable information. Thus, we leverage the affinity matrix $A^{(m)}$, which is supervised by clean samples, to constrain the gaze feature affinity matrix $A^{(f)}$, thereby reducing dependence on low-quality labels and transferring clean supervisory signals to noisy samples:

$$A_{i,j}^{(f)} = \frac{f_i^{\mathcal{N}} \cdot f_j^{\mathcal{C}}}{\|f_i^{\mathcal{N}}\| \|f_j^{\mathcal{C}}\|}, \tag{10}$$

$$\mathcal{L}_{\text{align}}^{\mathcal{N}} = -\frac{1}{B_N} \sum_{x_i \in \mathcal{D}_{\mathcal{S}}^{\mathcal{N}}} \frac{A_{i,:}^{(f)} \cdot A_{i,:}^{(m)}}{\|A_{i,:}^{(f)}\| \|A_{i,:}^{(m)}\|}, \tag{11}$$

where $f_i^{\mathcal{N}}$ and $f_j^{\mathcal{C}}$ denote f from $\mathcal{D}_{\mathcal{S}}^{\mathcal{N}}$ and $\mathcal{D}_{\mathcal{S}}^{\mathcal{C}}$ separately, B_N is the batch size of $\mathcal{D}_{\mathcal{S}}^{\mathcal{N}}$. $A_{i,:}^{(f)}$ and $A_{i,:}^{(m)}$ denote the feature and manifold affinities between x_i and clean samples in the batch respectively. Unlike Eq. 9, which adopts a relatively rigid MAE loss to precisely align the affinity matrix $A^{(m)}$, Eq. 11 relaxes the constraint through row-wise cosine similarity. This soft alignment focuses on refining the pairwise relationships among features while maintaining their separability in the high-dimensional space.

From the perspective of noisy label learning, our proposed noise-robust regularization strategy makes full use of clean samples and performs feature-level semantic correction on noisy samples, thereby encouraging the model to learn domain-invariant features, which can enhance robustness to unseen domains.

3.5. Overall

In summary, our final objective loss function is:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{gaze}} + \mathcal{L}_{\text{align}}^{\mathcal{C}} + \lambda \mathcal{L}_{\text{align}}^{\mathcal{N}}, \tag{12}$$

where λ is set to 0.1. During inference in unseen domains, SeeTN utilizes solely the backbone to extract features, which are then fed into the regression head to predict the final gaze angles.

4. Experiments

4.1. Dataset Preparation

We conduct experiments on four widely used gaze estimation datasets: ETH-XGaze ($\mathcal{D}_E$) [34], Gaze360 ($\mathcal{D}_G$) [14], MPIIFaceGaze ($\mathcal{D}_M$) [32], and EyeDiap ($\mathcal{D}_D$) [11]. **ETH-XGaze**: The images are captured by high-resolution cameras in a laboratory environment, covering a wide gaze range. **Gaze360**: The images are captured by a 360° camera in streets. We use 84,902 images with frontal faces in our experiments. **MPIIFaceGaze**: The images are captured by web camera of laptop computers. **EyeDiap**: The images are captured under laboratory environment with screen and floating targets. For a fair comparison, we preprocess data following the same procedures as previous studies [10], and report performance by using $\mathcal{D}_E$ and $\mathcal{D}_G$ as the source domains separately, with the remaining datasets serving as the target domains.

4.2. Implementation Details

SeeTN is implemented in PyTorch with ResNet-18 [13] and ResNet-50 [13] as the backbone. The model is trained using the Adam optimizer with a learning rate of 0.0001 for 100 epochs on Gaze360 and for 10 epochs on ETH-XGaze. During the early stage of training, we warm up the model for 10 epochs on Gaze360 and for 2 epochs on ETH-XGaze to obtain a reliable indicator η for partitioning the dataset. Besides, we set the number of prototypes to $K = 12$. In all the settings, batch sizes are set to 128.

4.3. Quantitative Evaluation

4.3.1. Effectiveness of SeeTN

We first evaluate SeeTN with different backbones to verify its effectiveness in mitigating the effects of noisy labels

Table 1. Performance comparison of SeeTN with different backbones (ResNet-18 and ResNet-50) under four cross-domain gaze estimation settings and within-domain settings. Results are shown as mean angular errors (°), with red values denoting relative improvement over the baseline. Lower values indicate better performance.

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	8.07	8.78	7.94	8.73	4.64	11.14
ResNet-18+SeeTN	**6.58** ↓18.5%	**7.18** ↓18.2%	**6.57** ↓17.2%	**7.57** ↓13.3%	**4.40** ↓5.1%	**10.73** ↓3.7%
ResNet-50	7.64	8.39	7.68	8.65	4.32	10.78
ResNet-50+SeeTN	**6.31** ↓17.4%	**6.84** ↓18.4%	**6.75** ↓12.1%	**7.42** ↓14.2%	**4.16** ↓3.7%	**10.45** ↓3.1%

Table 2. Comparison of domain generalization performance between SeeTN and state-of-the-art gaze estimation methods. $*$ indicates methods with ResNet-50 backbone.

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 [33]	12.35	30.15	11.13	14.42
ADL [23]	7.23	8.02	11.36	11.86
CA-Net [8]	–	–	27.13	31.41
LatentGaze [17]	7.98	9.81	–	–
PureGaze [9]	7.08*	7.48*	9.28	9.32
GazeCF [29]	6.50	7.44	7.55	9.00
CGaG [27]	6.47	7.03	7.50	8.67
AGG [2]	7.10	7.07	7.87	7.93
AGG* [2]	<u>5.91</u>	**6.75**	9.20	11.36
FSCI [19]	**5.79**	6.96	7.06	7.99
ResNet18+SeeTN	6.58	7.18	**6.57**	<u>7.57</u>
ResNet50+SeeTN*	6.31	<u>6.84</u>	<u>6.75</u>	**7.42**

Table 3. Comparison of domain generalization performance between SeeTN and representative noisy-label learning methods.

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
Baseline	8.07	8.78	7.94	8.73
DivideMix [18]	9.72	20.09	10.46	12.30
SUGE [24]	10.00	8.74	7.04	8.32
SeeTN	**6.58**	**7.18**	**6.57**	**7.57**

and enhancing the model's domain generalization capability. As shown in Tab. 1, we conduct experiments in four cross-domain and two within-domain settings. Specifically, when integrated with ResNet-18 and ResNet-50, SeeTN consistently improves the performance across all scenarios. SeeTN achieves improvements of 18.5% on $\mathcal{D}_E \rightarrow \mathcal{D}_M$ and 17.2% on $\mathcal{D}_G \rightarrow \mathcal{D}_M$, without using any target-domain data. Moreover, many generalization approaches [2] improve performance on unseen domains at the cost of degraded within-domain accuracy. In contrast, even under the within-domain settings $\mathcal{D}_E^*$ and $\mathcal{D}_G^*$, SeeTN still brings a 3–5% reduction in error. These results sufficiently demonstrate that SeeTN effectively alleviates the negative influence of noisy labels while substantially improving the model's robustness and generalization ability across different domains.

4.3.2. Comparison with Domain Generalization Methods

We compare the domain generalization ability of the proposed SeeTN framework with several recent gaze estimation methods [2, 8, 9, 17, 19, 23, 27, 29, 33]. As shown in Tab. 2, when trained on $\mathcal{D}_G$ as the source domain, SeeTN significantly outperforms other state-of-the-art methods, surpassing AGG [2] by 2.63 and FSCI [19] by 0.49. In contrast, when trained on $\mathcal{D}_E$ as the source domain, its performance is slightly inferior to that of AGG* and FSCI. One possible reason, as also discussed in AGG [2], is the instability of the baseline model under cross-domain settings. Another reason is that $\mathcal{D}_G$ is more noisy than $\mathcal{D}_E$, which further indicates that the performance gains of SeeTN mainly stem from its explicit handling of noisy labels. More analysis of SeeTN's ability to cope with noisy labels is provided in the supplementary material.

4.3.3. Comparison with Noisy Label Learning Methods

Since SeeTN improves domain generalization from the perspective of noisy labels, we compare it with DivideMix [18] and SUGE [24] under cross-domain settings. DivideMix is a representative method in LNL, and extensive empirical evidence has verified its effectiveness in handling label noise. However, as shown in Tab. 3, DivideMix performs poorly in gaze estimation and even underperforms the backbone, since it leverages MixMatch, a semi-supervised learning strategy designed for classification tasks, which is not well suited for regression problems. In gaze estimation, SUGE focuses on suppressing noisy labels in within-domain settings. We implement SUGE for cross-domain settings. It can be observed that SUGE achieves a slight improvement on $\mathcal{D}_G$, but its performance on $\mathcal{D}_E$ is unsatisfactory. SeeTN achieves significantly better performance than two previous methods, demonstrating the adaptability of our approach compared with existing LNL methods for domain generalization in gaze estimation.

Table 4. Ablation study for the effectiveness of each loss function on cross-domain settings.

L_{gaze}	$L^{\mathcal{C}}_{\text{align}}$	$L^{\mathcal{N}}_{\text{align}}$	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
✓			8.07	8.78	7.94	8.73
✓	✓		7.14	7.95	7.26	8.11
✓	✓	✓	**6.58**	**7.18**	**6.57**	**7.57**

Table 5. The ablation of different parameters and indicators on cross-domain scenarios.

Params.	Value	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
K	8	7.37	7.32	6.90	**7.36**
	12	**6.58**	**7.18**	6.57	7.57
	16	7.59	7.83	**6.44**	8.25
$t\%$	5	**6.58**	**7.18**	6.85	7.93
	10	7.25	7.42	**6.57**	**7.57**
	20	7.63	7.95	6.97	8.26
Indicator	L1	7.73	7.64	7.25	7.91
	η	**6.58**	**7.18**	**6.57**	**7.57**

4.4. Ablation Study

Effect of Loss Functions. We conduct ablation experiments to verify the effectiveness of each loss function designed in SeeTN. As shown in Tab. 4, removing $\mathcal{L}^{\mathcal{N}}_{\text{align}}$ leads to a significant performance reduction of 0.56° ($\mathcal{D}_E \rightarrow \mathcal{D}_M$), 0.77° ($\mathcal{D}_E \rightarrow \mathcal{D}_D$), 0.69° ($\mathcal{D}_G \rightarrow \mathcal{D}_M$) and 0.54° ($\mathcal{D}_G \rightarrow \mathcal{D}_D$). Removing $\mathcal{L}^{\mathcal{C}}_{\text{align}}$ results in a further performance decline. It can be concluded that both $\mathcal{L}^{\mathcal{C}}_{\text{align}}$ and $\mathcal{L}^{\mathcal{N}}_{\text{align}}$ improve the performance on all the cross-domain settings compared with the baseline, showing the significance of mitigating and utilizing noisy labels in domain generalization for gaze estimation.

Number of Prototypes μ (K). To explore the effect of the number of μ on SeeTN, we conduct experiments with different μ values under the cross-domain settings. As shown in Tab. 5, the performance shows a decreasing trend as K increases. We infer that, due to the lack of explicit semantic distinction among prototypes in regression tasks, a larger K makes the prototypes μ more likely to be affected by within-domain styles, leading to a decline in performance. Overall, $K = 12$ is an appropriate choice for the source domain.

Ratio of Divided Noisy Samples ($t\%$). In SeeTN, we regard the top $t\%$ of samples with larger η values as noisy samples. To explore the sensitivity of our method to t, we conduct experiments with various values of t, as shown in Tab. 5. In $\mathcal{D}_E$, we select $t\% = 5\%$ since the dataset is collected in a laboratory environment with lower noise ratio. In $\mathcal{D}_G$, we select $t\% = 10\%$ for Gaze360. As t increases, the performance shows a slight decrease, but it still remains considerably higher than that of the baseline.

Different Indicators. For gaze estimation tasks, we specially design a novel indicator η to distinguish noisy samples by pairwise relations among samples. To verify the effectiveness of η, we compare different indicators on cross-domain settings. Specifically, based on the small-loss criterion in noisy label learning, we replace η with L1 loss to identify noisy samples. As shown in Tab. 5, SeeTN with η achieves better performance than the L1 loss, showing the reliability of the proposed noisy detector. More ablations are provided in the supplementary material.

4.5. Qualitative Evaluation

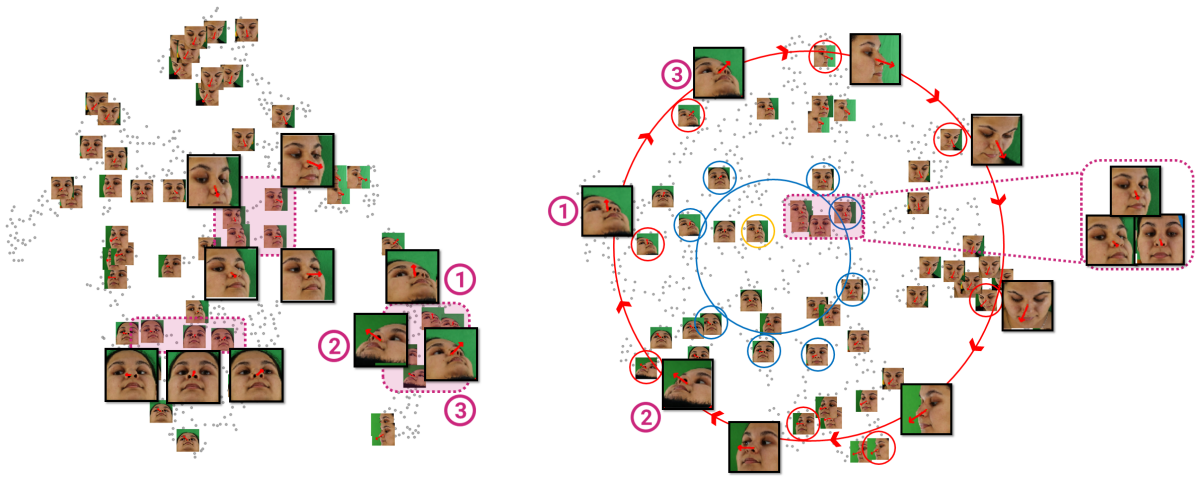
4.5.1. Semantic Feature Visualization

To qualitatively evaluate the effect of SeeTN on the feature space, we randomly select 800 images from one subject and visualize their feature representations under the baseline and our SeeTN using t-SNE [20].

As shown in the pink dashed regions of Fig. 3(a), it is evident that the baseline clusters the features mainly according to head pose rather than the more task-relevant gaze direction, which indicates that the model has not truly learned gaze-related semantics, and thus its cross-domain performance obviously decreases when the domain style changes.

Encouragingly, as shown in Fig. 3(b), the features learned by our SeeTN are strongly associated with the gaze direction rather than head pose, as illustrated by the pink dashed regions. From the visualization results, we observe that the feature distribution learned by SeeTN exhibits the following three properties:

- The clustering of sample features is mainly based on the gaze direction rather than the head pose. For instance, in the baseline visualization, three samples clustered in the lower-right corner share similar head poses, whereas in SeeTN, they are dispersed to different positions in the feature space according to their respective gaze directions.
- The labels corresponding to the sample features change continuously in the feature space, which can be illustrated by the samples located on the red and blue circles in Fig. 3(b). It can be observed that, in the t-SNE visualization of SeeTN, the gaze directions corresponding to these samples change continuously in a clockwise manner, demonstrating that our approach effectively preserves the inherent continuity of labels in gaze regression tasks.
- The deviation of the sample gaze direction from the frontal direction increases radially outward. For example, the gaze directions at corresponding radial positions on the blue and red circles exhibit nearly identical in the images, whereas the label arrows become progressively longer toward the periphery. The longer the arrow is, the larger the deviation angle is between the gaze direction and the frontal direction. This further indicates that the features learned by our SeeTN exhibit semantic continu-



(a) Feature Visualization of Baseline

(b) Feature Visualization of SeeTN

Figure 3. The qualitative results of visualizing the features learned by the backbone and SeeTN on the ETH-XGaze dataset via t-SNE. The red arrows in the images denote the label directions of the samples.

ity with respect to the gaze labels.

The aforementioned observations and analyses clearly validate the consistency between the motivation, design methodology, and experimental results of the proposed SeeTN. By enforcing that similar features exhibit similar gaze directions within a semantic manifold space, we embed semantic information into the feature representation and enhance the association between features and gaze direction. This design reduces the model's reliance on domain-specific style information, thereby improving its generalization ability across domains.

4.5.2. Visualization of Noisy Sample Detection

To intuitively illustrate how the indicator η evaluates the samples, we visualize a subset of samples with the smallest η values and those with the largest η values. As shown in Fig. 4, for samples with low η values, the labels are accurate, and the model's predicted directions nearly coincide with the gaze labels. For samples with large η values, their labels are often incorrect, while the model can still accurately predict their gaze directions. These results clearly demonstrate that the indicator η can effectively assess whether a sample contains noisy labels, and also show that our model is robust to label noise. Additional visualization results can be found in the supplementary material.

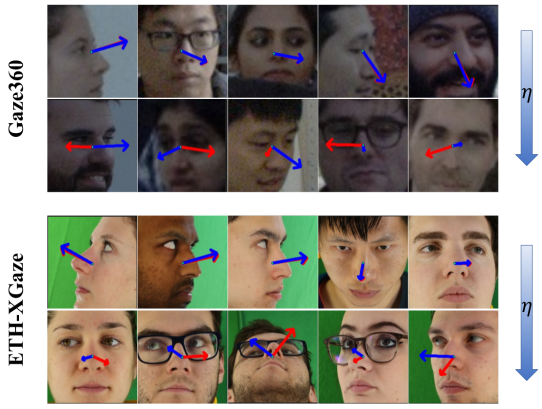


Figure 4. Sample Visualization with the smallest and largest η values in Gaze360 and ETH-XGaze. The red and blue arrows indicate the ground-truth directions and the model predictions respectively.

5. Conclusion and Discussion

In this paper, we propose an effective framework named SeeTN to improve domain generalization performance in gaze estimation from a novel perspective of mitigating the negative impact of noisy labels. We first design a semantic manifold to preserve relationships between features and labels. We then measure feature-label affinity consistency to identify noisy samples, and introduce a noise-robust regularization in the semantic manifold to transfer gaze-related information from clean to noisy samples. Compared with existing methods, SeeTN achieves superior domain generalization performance without sacrificing within-domain accuracy, demonstrating the effectiveness of handling noisy samples in cross-domain gaze estimation task.

Limitation. While SeeTN has achieved significant improvement for various cross-domain settings, one limitation is that the strategy used to distinguish between clean and noisy samples is relatively simple. We will investigate more effective approaches that can accurately divide the noisy subset by combining noisy label learning methods with the semantic manifold proposed in this paper.

Acknowledgements

This work is supported by National Natural Science Foundation of China (62271042, 62376021) and Beijing Natural Science Foundation (L259055).

References

[1] Ehsan Amid, Manfred KK Warmuth, Rohan Anil, and Tomer Koren. Robust bi-tempered logistic loss based on bregman divergences. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[2] Yiwei Bao and Feng Lu. From feature to gaze: A generalizable replacement of linear layer for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1409–1418, 2024. 2, 6

[3] 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. 1, 2

[4] Pradipta Biswas et al. Appearance-based gaze estimation using attention and difference mechanism. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3143–3152, 2021. 1

[5] 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

[6] Xin Cai, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Source-free adaptive gaze estimation by uncertainty reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22035–22045, 2023. 2

[7] 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

[8] 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. 6

[9] 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, 6

[10] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7509–7528, 2024. 5

[11] 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. 5

[12] Xian-Jin Gui, Wei Wang, and Zhang-Hao Tian. Towards understanding deep learning from noisy labels with small-loss criterion. *arXiv preprint arXiv:2106.09291*, 2021. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 5

[14] 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, 5

[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] 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

[17] 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. 6

[18] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020. 3, 6

[19] Ziyang Liang, Yiwei Bao, and Feng Lu. De-confounded gaze estimation. In *European Conference on Computer Vision*, pages 219–235. Springer, 2024. 6

[20] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008. 7

[21] 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. 1

[22] Yuanpeng Tu, Boshen Zhang, Yuxi Li, Liang Liu, Jian Li, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Cai Rong Zhao. Learning with noisy labels via self-supervised adversarial noisy masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16186–16195, 2023. 3

[23] 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, 6

[24] Shijing Wang and Yaping Huang. Suppressing uncertainty in gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5581–5589, 2024. 3, 6

[25] 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

[26] Qi Wei, Lei Feng, Haoliang Sun, Ren Wang, Chenhui Guo, and Yilong Yin. Fine-grained classification with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11651–11660, 2023. 3

[27] Lifan Xia, Yong Li, Xin Cai, Zhen Cui, Chunyan Xu, and Antoni B Chan. Collaborative contrastive learning for cross-domain gaze estimation. *Pattern Recognition*, 161:111244, 2025. 1, 2, 6

[28] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015. 3

[29] Mingjie Xu, Haofei Wang, and Feng Lu. Learning a generalized gaze estimator from gaze-consistent feature. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3027–3035, 2023. 2, 6

[30] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International conference on machine learning*, pages 7164–7173. PMLR, 2019. 3

[31] 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. 1

[32] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017. 5

[33] 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. 6

[34] 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. 5

[35] Ziyi Zhang, Weikai Chen, Chaowei Fang, Zhen Li, Lechao Chen, Liang Lin, and Guanbin Li. Rankmatch: Fostering confidence and consistency in learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1644–1654, 2023. 3