

Suppressing Uncertainty in Gaze Estimation

Shijing Wang, Yaping Huang*

Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, China
{shijingwang, yphuang}@bjtu.edu.cn

Abstract

Uncertainty in gaze estimation manifests in two aspects: 1) low-quality images caused by occlusion, blurriness, inconsistent eye movements, or even non-face images; 2) incorrect labels resulting from the misalignment between the labeled and actual gaze points during the annotation process. Allowing these uncertainties to participate in training hinders the improvement of gaze estimation. To tackle these challenges, in this paper, we propose an effective solution, named Suppressing Uncertainty in Gaze Estimation (SUGE), which introduces a novel triplet-label consistency measurement to estimate and reduce the uncertainties. Specifically, for each training sample, we propose to estimate a novel “neighboring label” calculated by a linearly weighted projection from the neighbors to capture the similarity relationship between image features and their corresponding labels, which can be incorporated with the predicted pseudo label and ground-truth label for uncertainty estimation. By modeling such triplet-label consistency, we can measure the qualities of both images and labels, and further largely reduce the negative effects of unqualified images and wrong labels through our designed sample weighting and label correction strategies. Experimental results on the gaze estimation benchmarks indicate that our proposed SUGE achieves state-of-the-art performance.

Introduction

Gaze estimation is a crucial task in computer vision that aims to accurately determine the direction of a person’s gaze based on visual cues. In recent years, gaze estimation has gained significant attention due to its wide-ranging applications in fields such as human-computer interaction (Majaranta and Bulling 2014) (Rahal and Fiedler 2019), virtual reality (Patney et al. 2016) (Kim et al. 2019), and assistive technology (Jiang and Zhao 2017) (Liu, Li, and Yi 2016) (Dias et al. 2020). Benefiting from the deep learning techniques and large-scale training data, appearance-based gaze estimation has made rapid progress and achieved promising results.

Recent works are mainly dedicated to developing advanced networks (Zhang et al. 2017a) (Kellnhofer et al. 2019) (Cheng et al. 2020) (Cheng and Lu 2022) to extract

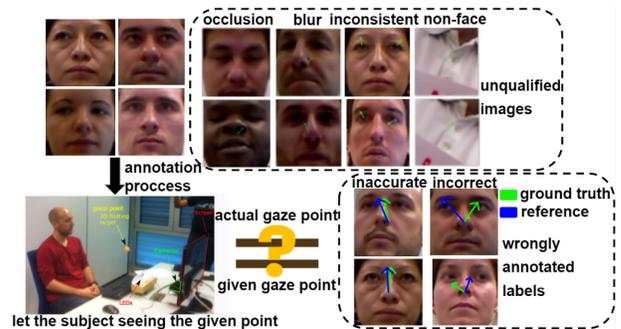


Figure 1: Illustration of uncertainties in gaze estimation using the EyeDiap dataset as an example. The upper half of the figure reflects unqualified images, where the images in right side are extremely difficult for machines and even human. These images are better to be suppressed in training. The lower half of the figure reflects wrongly annotated labels, where the left side represents the common data annotation process. Due to the challenge of achieving perfect alignment between the actual gaze points and the given gaze points, inaccurate and incorrect labels exist in the datasets, which should be rectified.

distinctive gaze features. However, the success of deep models heavily relies on sufficient amount of data with accurate ground truth labels. Unfortunately, it is very challenging for humans to provide consistent and precise annotations for gaze estimation task, especially in the complex natural scenes. Common gaze estimation datasets (Funes Mora, Monay, and Odobez 2014) (Zhang et al. 2017b) (Kellnhofer et al. 2019) (Zhang et al. 2020b) usually obtain annotations by requiring subjects to fixate on given points during data collection. But this annotation strategy is based on the assumption of the perfect alignment between the actual gaze points and the provided points, which is practically impossible to achieve.

As shown in Fig. 1, many captured gaze images suffer from serious quality degradation due to eyelid occlusion, blurriness of the eyes, inconsistent eye movements, which may lead to some unqualified eye images, or even background images are involved in the collecting process. On the other hand, there are also a significant number of visibly in-

*Corresponding author

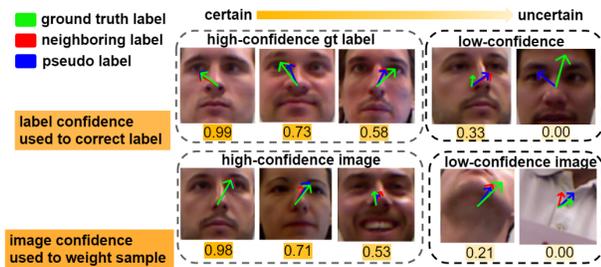


Figure 2: Visualization results of varying image and label confidences using samples from folds 0, 1, 3 of the EyeDiap dataset as the training set in the subsequent epoch after the warm-up phase. These results showcase the effectiveness of the two uncertainty metrics we design.

correct labels presented in the commonly used datasets. Allowing these unqualified images and incorrect ground truth labels to be included in the training process may result in overfitting, which hinders the model from learning the discriminative features for accurate gaze estimation. Generally, although many efforts have been made for getting precise annotations, the noisy data and labels are inevitably introduced, which is neglected in the previous works.

To address the above-mentioned issues, in this paper, we propose a novel solution, named “Suppressing Uncertainty in Gaze Estimation” (SUGE), which brings a new perspective of uncertainty estimation for gaze estimation task. The key issue of our work is how to effectively estimate the uncertainty for measuring the quality of images and labels, and meanwhile alleviating the negative effects of them. To achieve this goal, we propose a novel triplet-label consistency method to estimate the uncertainty, where “neighboring label” is proposed by computing a weighted average of labels from neighboring image features, and coupled with the predicted labels and ground truth labels to calculate two uncertainty metrics of each training sample by using Gaussian Mixture Model (GMM). As shown in Fig. 2, by modeling two uncertainty metrics, we can obtain two confidences: image confidence and label confidence, where the former reflects the quality of the images and the latter measures the correctness of the annotations. Afterwards, we utilize the estimated confidences for guiding the further training process: the image confidence is used to weight the training sample, and the label confidence is referred when performing label correction with the predicted pseudo-label and neighboring label. The effectiveness of our proposed SUGE is comprehensively evaluated on popular gaze estimation benchmarks.

In summary, our contribution is three-fold:

- We address the gaze estimation task from a new perspective of uncertainty estimation to mitigate the effect of low-quality images and incorrect annotations, which commonly exists in real-world gaze estimation datasets but is ignored in previous works.
- We propose a novel triplet-label consistency measurement to estimate the uncertainty, where a novel “neighboring label” representing the local consistency is coupled with the predicted pseudo label and ground truth la-

bel to assess the uncertainty of samples. Then the produced uncertainty is further utilized for better training by the proposed label correction and sample weighting strategies.

- We conduct comprehensive experiments on real-world gaze estimation benchmarks, and the experimental results demonstrate that our method achieves state-of-the-art performance.

Related Works

Gaze Estimation

Gaze estimation methods are categorized into model-based and appearance-based approaches (Hansen and Ji 2009). Model-based techniques (Zhu and Ji 2007; Valenti, Sebe, and Gevers 2011; Alberto Funes Mora and Odobez 2014) utilize 3D eye models with specialized equipment in controlled settings, while appearance-based methods (Cheng et al. 2021) employ machine learning to map images to gaze directions, gaining popularity for their adaptability.

Data annotation poses a critical challenge for appearance-based gaze estimation. Two primary methods are prevalent. The first involves model-based eye-tracking devices, such as desktop eye trackers (Park et al. 2020) or head-mounted ones (Fischer, Chang, and Demiris 2018), while the former is restricted to specific distances and head poses, and the latter severely occludes eye appearance. The second method employs fixation-based annotation scheme, allowing subjects to focus on specific points of interest (Smith et al. 2013; Sugano, Matsushita, and Sato 2014; Funes Mora, Monay, and Odobez 2014; Zhang et al. 2017b; Kellnhofer et al. 2019; Zhang et al. 2020b). This approach remains flexible, unaffected by appearance changes, making it cost-effective and widely adopted for creating gaze estimation datasets.

Despite great efforts have been made in building gaze estimation datasets, the complexity of generating gaze annotations inevitably introduces low-quality data and labels. These issues are often overlooked in existing works and have become significant bottlenecks hindering the development of gaze estimation algorithms. Our work represents the first attempt to address the problem of annotation quality from a novel perspective of uncertainty estimation.

Uncertainty Estimation in Computer Vision

Uncertainty estimation is crucial for capturing the randomness in the learning process across computer vision tasks, such as facial expression recognition (Wang et al. 2020), saliency detection (Zhang et al. 2020a), and edge detection (Zhou et al. 2023). For gaze estimation, handling data uncertainty is crucial. Approaches like (Kellnhofer et al. 2019; Dias et al. 2020) incorporate uncertainty prediction heads at the output layer to handle high uncertainty data. Nonaka et al. (Nonaka, Nobuhara, and Nishino 2022) focus on multi-source input uncertainty by adding prediction heads for each input’s features. Recent work (Cai et al. 2023) tackles cross-domain adaptation by enhancing image quality to mitigate image uncertainty and reducing prediction variance to manage model uncertainty.

In contrast, our approach uniquely addresses low-quality samples and inaccurately annotated labels in gaze estimation datasets. We propose a novel method to assess label consistency, estimating uncertainty in both image and label space. This tackles low-quality images and rectifies inaccurate labels, resulting in significant performance improvements.

Method

To mitigate the adverse impact of data uncertainty in both image space and label space of gaze estimation, we propose a novel Suppressing Uncertainty in Gaze Estimation (SUGE) method. In this section, we will begin by giving an overview of SUGE and subsequently introduce its three modules and co-training strategy in detail.

Overview of SUGE

In order to capture the data uncertainty both in image and label space and further guide the following training process, we develop a novel triplet-label consistency for uncertainty estimation. The intuition behind our design is based on two foundations: 1) The discrepancy between ground truth labels and pseudo labels for a given sample indicates its training difficulty and modeling complexity; 2) The discrepancy between neighboring labels and the other two types of labels reflects local relationship and smoothness between image features and their corresponding labels. By modeling the described triplet-label consistency, we can purify the quality of samples and their corresponding labels, which effectively reduces overfitting risks and enhances overall generalization.

To realize our idea, as depicted in the Fig. 3, our proposed SUGE is composed of three main modules: the neighboring labeling module, the uncertainty estimation module, as well as the label correction and sample weighting module. Specifically, for each input image, the encoder first extracts features, and the neighboring labeling module calculates the neighboring labels for the sample based on the ground truth labels of its neighbors in feature space. The uncertainty estimation module designs the uncertainty metrics of label and image by considering the consistency among neighboring labels, pseudo labels, and ground truth labels. Two uncertainty metrics are then input into the Gaussian Mixture Model (GMM) (Permuter, Francos, and Jermyn 2006) to calculate confidence scores. Finally, the label correction and sample weighting module utilizes the confidence scores to perform label correction and sample weighting for training.

Furthermore, to avoid overconfidence and cumulative errors during self-training of a single network on its own generated image and label confidence, we adopt the approach from the paper (Friend, Reising, and Cook 1993) (Li, Socher, and Hoi 2020) and train two networks simultaneously. Both networks contribute to generating corrected labels and sample weights for each other. The final algorithm is presented in Algorithm 1.

Neighboring Labeling Module

In this module, we aim to capture the local similarity relationship between image features and their corresponding labels by computing neighboring labels. Formally, given a

Algorithm 1: SUGE

Input: two encoder and fully connected parameters $(\mathcal{E}^{(1)}, f^{(1)})$ and $(\mathcal{E}^{(2)}, f^{(2)})$, training data (X, Y)
Param.: small constant for denominator ϵ , number of neighbors K , confidence threshold τ

- 1: $(\mathcal{E}^{(1)}, f^{(1)}) \leftarrow \text{WarmUp}(X, Y, (\mathcal{E}^{(1)}, f^{(1)}))$
- 2: $(\mathcal{E}^{(2)}, f^{(2)}) \leftarrow \text{WarmUp}(X, Y, (\mathcal{E}^{(2)}, f^{(2)}))$
- 3: **while** $e < \text{MaxEpoch}$ **do**
- 4: **for** $k = 1, 2$ **do**
- 5: $\hat{Y}^p \leftarrow f^{(k)}(\mathcal{E}^{(k)}(X))$
- 6: // Neighbor labeling module
- 7: $X_{ij} \leftarrow \text{K-NN}(\mathcal{E}^{(k)}(X_i), K)$, for $i = 1, \dots, N$
- 8: Get the reconstruction weight A by Equation 2
- 9: $\hat{Y}_i^n \leftarrow \sum_{j=1}^K Y_{ij} * A_{ij}$, for $i = 1, \dots, N$
- 10: // Uncertainty estimation module
- 11: $D^{pg}, D^{pn}, D^{ng} \leftarrow \text{AngularDis}(Y, \hat{Y}^n, \hat{Y}^p)$
- 12: $\text{TupleMD} \leftarrow \frac{\min(D^{pg}, D^{ng})}{D^{pn} + \epsilon}$
- 13: $\text{TripleMD} \leftarrow \min(D^{pg}, D^{pn}, D^{ng})$
- 14: $\Gamma^{\text{label}} \leftarrow \text{GMM}(\text{TupleMD})$
- 15: $\Gamma^{\text{image}} \leftarrow \text{GMM}(\text{TripleMD})$
- 16: // Label correction and sample weighting
- 17: $\Gamma^{\text{label}} \leftarrow \text{Truncate}(\Gamma^{\text{label}}, \tau, 0)$
- 18: Calculate $\hat{Y}^{(k)}$ using Γ^{label} by Equation 14
- 19: $\hat{\Gamma}^{(k)} \leftarrow \Gamma^{\text{label}}$
- 20: $\Gamma^{\text{image}} \leftarrow \text{Truncate}(\Gamma^{\text{image}}, \tau, 0)$
- 21: $\hat{W}^{(k)} \leftarrow \Gamma^{\text{image}}$
- 22: **end for**
- 23: If $\hat{\Gamma}_i^{(1)} = 0$, set $\hat{Y}_i^{(1)} = \frac{\hat{Y}_i^{(1)} + \hat{Y}_i^{(2)}}{2}$, for $i = 1, \dots, N$
- 24: If $\hat{\Gamma}_i^{(2)} = 0$, set $\hat{Y}_i^{(2)} = \frac{\hat{Y}_i^{(1)} + \hat{Y}_i^{(2)}}{2}$, for $i = 1, \dots, N$
- 25: $(\mathcal{E}^{(1)}, f^{(1)}) \leftarrow \text{Train}(X, \hat{Y}^{(2)}, (\mathcal{E}^{(1)}, f^{(1)}), \hat{W}^{(2)})$
- 26: $(\mathcal{E}^{(2)}, f^{(2)}) \leftarrow \text{Train}(X, \hat{Y}^{(1)}, (\mathcal{E}^{(2)}, f^{(2)}), \hat{W}^{(1)})$
- 27: **end while**

training dataset $\mathcal{D} = \{X, Y\} = \{(X_1, Y_1), \dots, (X_N, Y_N)\}$, where X_i is i^{th} input image and $Y_i = (Y_{i,\psi}, Y_{i,\theta})^T$ is the ground truth 2D gaze angle vector (yaw and pitch), we use a K nearest-neighbor (K-NN) algorithm to generate the j^{th} neighbor of X_i denoted by X_{ij} , $j = 1, \dots, K$ where the neighboring samples are identified based on the distance between samples in the feature space $\mathcal{E}(X)$, and are restricted to the same person ID as sample X_i .

Next, we find the optimal reconstruction weight A_i for each sample X_i with respect to its neighboring samples X_{ij} , $j = 1, \dots, K$. This is achieved by solving the following optimization problem:

$$\begin{aligned} & \underset{A_i}{\text{minimize}} \|\mathcal{E}(X_i) - \sum_{j=1}^K (A_{ij} \mathcal{E}(X_{ij}))\|^2 + \lambda \|A_i\|^2, \\ & \text{subject to: } \sum_{j=1}^K A_{ij} = 1. \end{aligned} \quad (1)$$

The closed-form solution for the optimal reconstruction

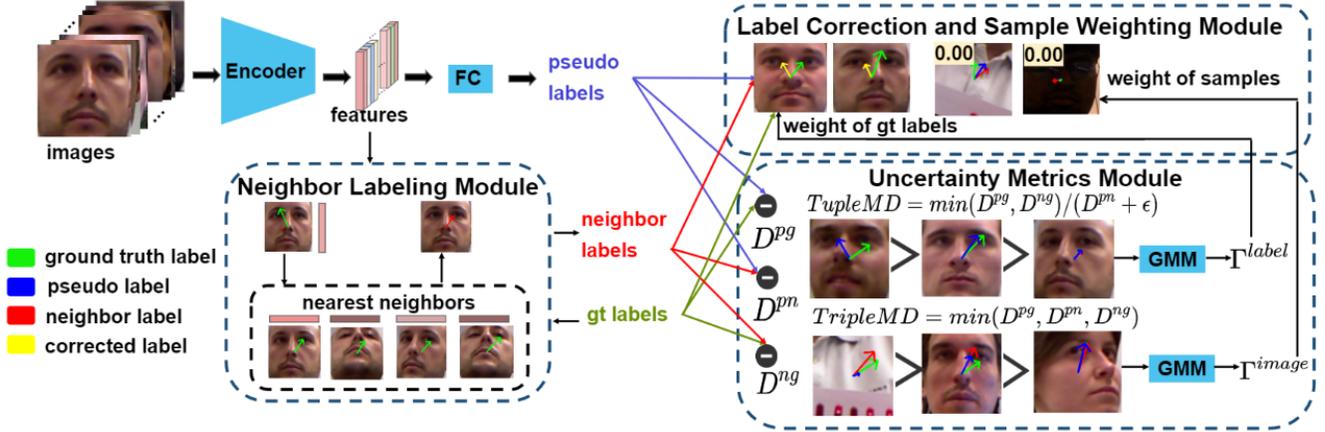


Figure 3: The pipeline of SUGE method. Initially, input images undergo feature extraction through the encoder, and a fully connected layer generates pseudo labels. The neighboring labeling module then employs a nearest-neighbor algorithm to find feature neighbors for each image and calculates the neighboring label by weighted averaging of the ground truth labels from its neighbors. Next, the Uncertainty Metrics module comes into play, computing Tuple Minimum Discrepancy and Triple Minimum Discrepancy by measuring the consistency among pseudo labels, ground truth labels, and neighboring labels. These uncertainties metrics are further input into a Gaussian Mixture Model, which yields two confidence scores: label confidence and image confidence. In the Label Correction and Sample Weighting module, label confidence is employed to perform weighted calculations on the ground truth labels, pseudo labels, and neighboring labels, resulting in corrected labels. Additionally, the sample weight is determined based on the image confidence and can further be used to guide the training process.

weights is given by:

$$A_i = \frac{(S_i + \lambda I)^{-1} \mathbf{1}_K}{\mathbf{1}_K^T (S_i + \lambda I)^{-1} \mathbf{1}_K}, \quad (2)$$

where

$$\begin{aligned} \Delta(X_i) &= [\mathcal{E}(X_i) - \mathcal{E}(X_{i1}), \dots, \mathcal{E}(X_i) - \mathcal{E}(X_{iK})], \\ S_i &= \Delta(X_i) \Delta(X_i)^T, \end{aligned} \quad (3)$$

I denotes the identity matrix, and $\mathbf{1}_K$ is a column vector containing K elements with all elements equal to 1.

Finally, we compute the new neighboring labels \hat{Y}_i^N by element-wise multiplying the original labels Y_{ij} with the optimal reconstruction weights A_{ij} :

$$\hat{Y}_i^N = \sum_{j=1}^K Y_{ij} * A_{ij}. \quad (4)$$

Through above process, we obtain a new set of labels $\hat{Y}^n = \{\hat{Y}_1^n, \dots, \hat{Y}_N^n\}$ called neighboring labels that capture the local similarity and smoothness between images and labels in the feature space.

Uncertainty Estimation Module

This module aims to assess the degree of uncertainty caused by low-quality labels and images by designing suitable metrics based on the consistency among the predicted pseudo label $\hat{Y}_i^p = f(\mathcal{E}(X_i))$, the neighboring label \hat{Y}_i^n , and the ground truth label Y_i for a given sample X_i . We focus on the angular differences among these label types, denoting them as D_i^{pg} , D_i^{pn} , and D_i^{ng} for measuring the triplet-label

consistency. Here taking D_i^{pg} as an example, D_i^{pn} and D_i^{ng} can be given following the similar calculation.

Firstly, we convert the labels from polar coordinate system to 3D Cartesian coordinate system using the `gazeto3d` function:

$$\text{gazeto3d}(\hat{Y}_i^p) = \begin{bmatrix} -\cos(\hat{Y}_{i,\theta}^p) \cdot \sin(\hat{Y}_{i,\psi}^p) \\ -\sin(\hat{Y}_{i,\theta}^p) \\ -\cos(\hat{Y}_{i,\theta}^p) \cdot \cos(\hat{Y}_{i,\psi}^p) \end{bmatrix}, \quad (5)$$

$$\text{gazeto3d}(Y_i) = \begin{bmatrix} -\cos(Y_{i,\theta}) \cdot \sin(Y_{i,\psi}) \\ -\sin(Y_{i,\theta}) \\ -\cos(Y_{i,\theta}) \cdot \cos(Y_{i,\psi}) \end{bmatrix}. \quad (6)$$

Next, we compute the angular distance between the two 3D coordinates using the angular function:

$$D_i^{pg} = \arccos \left(\frac{\text{gazeto3d}(\hat{Y}_i^p) \cdot \text{gazeto3d}(Y_i)}{\|\text{gazeto3d}(\hat{Y}_i^p)\| \cdot \|\text{gazeto3d}(Y_i)\|} \right) \cdot \frac{180}{\pi}, \quad (7)$$

where $\|\cdot\|$ denotes the Euclidean norm.

To measure the degree of uncertainty related to the label quality, we propose **TupleMD** as follows:

$$\text{TupleMD}_i = \frac{\min(D_i^{pg}, D_i^{ng})}{D_i^{pn} + \epsilon}. \quad (8)$$

The numerator of **TupleMD** is the minimum difference between the GT label and the other two labels (the pseudo label and neighboring label), which means that a GT label is considered unreliable when it exhibits substantial discrepancies with both two other labels. The denominator of **TupleMD** is a scale factor defined by the distance between the pseudo

label and the neighboring label, reflecting the uncertainty degree they contribute to the numerator. In addition to D_i^{pn} , the denominator also includes a constant term ϵ , to ensure that uncertainty is not erroneously considered high when both the numerator and denominator are very small.

Furthermore, to represent the degree of uncertainty related to image quality, we propose **TripleMD** as follows:

$$TripleMD_i = \min(D_i^{pg}, D_i^{pn}, D_i^{ng}). \quad (9)$$

The motivation behind this design is that the proposed three labels can be viewed as the representations provided by three ‘experts’, each offering their findings for the given image. GT labels and pseudo labels reflect the difficulties in data collection and model training respectively, and neighboring labels represent the relationship between similar images. Therefore, if any two findings exhibit significant discrepancies for a particular image, we treat it as a low-quality image.

Finally, to determine the probabilities of the above two uncertainty metrics, we adopt the bimodal Gaussian Mixture Model (GMM) to partition the two uncertainty metrics *TupleMD* and *TripleMD* into reliable and unreliable components. For each sample X_i , we define the label confidences Γ_i^{label} as the posterior probability $p(g|TupleMD_i)$ and image confidence Γ_i^{image} as the posterior probability $p(g|TripleMD)$, where g represents the Gaussian component with the smaller mean, indicating the likelihood of the sample being considered reliable. This allows us to quantify the uncertainty and assess the quality of the labels and images in the dataset.

Label Correction and Sample Weighting Module

In this module, we generate corrected labels and sample weights based on the confidences of labels and images for better training. We observe that the majority of label and image confidences are close to 1, while only a small number of confidence scores are low, indicating severe issues. To address this, for each sample X_i with confidence Γ_i , we set a threshold and truncate low confidences below the threshold to zero:

$$\Gamma_i = \begin{cases} \Gamma_i & \text{if } \Gamma_i > \tau \\ 0 & \text{otherwise} \end{cases}, \quad (10)$$

where Γ_i represents the label or image confidence of X_i , and τ is the threshold value.

Label correction. With the updated label confidence Γ_i^{label} , we can correct the ground truth label Y_i by pseudo label \hat{Y}_i^p and neighboring label \hat{Y}_i^n . To further enhance the accuracy of the corrected label, we employ the widely used gaze estimation augmentation technique of horizontal flipping on the original images X_i , maintaining unchanged eye gaze pitch while inverting the yaw component. This yields the augmented pseudo label \hat{Y}_i^{pa} :

$$\hat{Y}_i^{fpa} = f(\mathcal{E}(\text{HorizontalFlip}(X_i))), \quad (11)$$

$$\hat{Y}_i^{pa} = (-\hat{Y}_{i,\psi}^{fpa}, \hat{Y}_{i,\theta}^{fpa})^T, \quad (12)$$

where \mathcal{E} represents the encoder and f represents the fully connected layer.

Next, we use the neighboring sample X_{ij} ($j = 1 \dots K$) and the corresponding reconstruction weight A_{ij} of sample X_i obtained from the neighboring labeling module to calculate pseudo neighboring labels \hat{Y}_i^{np} and pseudo augmented neighboring labels \hat{Y}_i^{npa} based on \hat{Y}_i^p and \hat{Y}_i^{pa} . The formulations are as follows:

$$\hat{Y}_i^{np} = \sum_{j=1}^K \hat{Y}_{ij}^p * A_{ij}, \hat{Y}_i^{npa} = \sum_{j=1}^K \hat{Y}_{ij}^{pa} * A_{ij}. \quad (13)$$

Finally, the corrected labels are computed as a combination of the ground truth labels and the generated pseudo labels:

$$\hat{Y}_i = \Gamma_i^{label} \cdot Y_i + (1 - \Gamma_i^{label}) \cdot \frac{1}{5} (\hat{Y}_i^n + \hat{Y}_i^p + \hat{Y}_i^{pa} + \hat{Y}_i^{np} + \hat{Y}_i^{npa}). \quad (14)$$

Sample weighting. Additionally, to reduce the effect of low-quality samples, we further exploit the image confidence as a weight $\hat{W}_i = \Gamma_i^{image}$ to guide the training process. The overall loss objective is as follows:

$$\mathcal{L} = \sum_i \hat{W}_i \cdot \|\hat{Y}_i - f(\mathcal{E}(X_i))\|_1. \quad (15)$$

Co-training Strategy

To further alleviate the data uncertainty both in image and label space, we follow (Friend, Reising, and Cook 1993) (Li, Socher, and Hoi 2020) to introduce co-training strategy for maintaining two networks simultaneously, where they exchange the corrected labels \hat{Y} and sample weights \hat{W} in each iteration to prevent excessive confidence in their self-evaluation. Furthermore, for labels with truncated confidence equaling to zero, we additionally correct them by averaging the corrected labels generated by both networks.

Experiment

Dataset

The experiments utilize four widely used gaze estimation datasets: EyeDiap (Funes Mora, Monay, and Odobez 2014), MPIIFaceGaze (Zhang et al. 2017b), Gaze360 (Kellnhofer et al. 2019), and ETH-XGaze (Zhang et al. 2020b) (solely utilized for pretraining the GazeTR model). For a fair comparison, the data partitioning and preprocessing techniques for these datasets are maintained consistently with prior studies, as outlined by Cheng et al. (Cheng et al. 2021).

Implementation Details

Since our approach primarily focuses on mitigating data uncertainty through label correction and weighting, regardless of specific network architectures, we directly adopt two representative state-of-the-art (SOTA) methods namely Gaze360 (Kellnhofer et al. 2019) and GazeTR (Cheng and Lu 2022) implemented by Cheng et al. (Cheng et al. 2021) as baselines in our subsequent experiments. We utilize the same network architecture and corresponding parameter settings as these methods. Only for addressing the challenge of high-dimensional feature clustering, we reduce the feature dimension to 16 at the final layer of the encoder.

Method	Backbone	Dataset		
		MPII	Gaze360	EyeDiap
Full Face	C	4.93	14.99	6.53
CA-Net	C	4.27	11.20	5.27
Gaze360	C	4.06	11.04	5.36
CADSE	T	4.04	10.70	5.25
GazeTR	T	4.00	10.62	5.17
Gaze360 [†]	C	4.17	10.78	5.46
GazeTR [†]	T	4.00	10.61	5.34
SUGE (Gaze360)	C	4.07	10.52	5.05
SUGE (GazeTR)	T	4.01	10.51	5.04

Table 1: Comparison of gaze estimation performance in terms of angle error ($^{\circ}$) on three datasets. [†]represents our re-implemented results. C and T represent CNN (pretrained by Imagenet) and Transformer (pretrained by ETH-XGaze) backbone respectively.

Regarding our method, we set ϵ to 1 to prevent excessive *TupleMD*. Additionally, for the K nearest neighbor algorithm, we choose the KD tree with $K = 4$. Moreover, we set the warm-up epochs to 10, and the thresholds τ for truncating label and image confidences are both set to 0.5.

Comparison with SOTA Gaze Estimation Methods

We compare our method with state-of-the-art gaze estimation methods, including classical CNN-based methods pretrained on ImageNet, such as FullFace (Zhang et al. 2017a), CA-Net (Cheng et al. 2020) and Gaze360 (Kellnhofer et al. 2019), as well as Transformer-based methods pretrained on ETH-XGaze, such as CADSE (O Oh, Chang, and Choi 2022) and GazeTR (Cheng and Lu 2022). Our method focuses on purifying the data and labels, which can be easily combined with other models, so we re-implement the leading open-source methods from both categories, namely Gaze360[†] and GazeTR[†], and verify the effectiveness of our proposed method by assessing their performance with and without our SUGE approach on the EyeDiap, MPIIFaceGaze, and Gaze360 datasets.

As shown in Table 1, compared with CNN-based Gaze360[†], our approach reduces angle errors by 0.1° , 0.26° , and 0.41° on MPII, Gaze360 and EyeDiap datasets. Compared with Transformer-based GazeTR[†], the angle errors remain consistent on MPII dataset, while they are notably reduced by 0.10° and 0.30° on Gaze360 and EyeDiap datasets. These results significantly suggest that our approach can improve the quality of datasets, thereby ultimately achieving the most up-to-date SOTA results in gaze estimation task.

The Effectiveness of Our Uncertainty Estimation and Label Correction Paradigm

The issue of low-quality labels in gaze estimation is analogous to the classic problem of noisy label learning. To demonstrate the superiority of our design, we adopt the popular noise label learning methods on the Gaze360 baseline model and compare them with our proposed strategy on

Gaze360 and EyeDiap datasets. It should be noted that most latest methods (Li et al. 2023) (Wei et al. 2023) are basically designed for classification problems and can not be suitable for our gaze regression task. Therefore, we carefully choose the most notable noise learning strategies: Co-Teaching (Han et al. 2018) and DivideMix (Li, Socher, and Hoi 2020) for verification. These two strategies can be readily adapted for regression tasks since they directly utilize the loss as the sample selection criterion. Remarkably, DivideMix still maintains its preeminence as a leading method in the most real-world benchmarks of learning from noisy labels, like Clothing 1M.

Method	Dataset	
	Gaze360	EyeDiap
Baseline	10.78	5.46
CoTeaching	10.59	5.18
DivideMix	10.66	5.16
SUGE	10.52	5.05

Table 2: Comparison in terms of angle error ($^{\circ}$) on two datasets with noise label learning methods.

The comparisons with other label noisy methods are summarized in Table 2. We can see that our method outperforms two excellent noise label methods in the gaze estimation task, demonstrating that our proposed uncertainty estimation can successfully access the quality of data and label and further effectively guide the following training process. Moreover, all noise label learning strategies, including our SUGE, surpass the baseline model, highlighting the importance of addressing low-quality labels in the gaze estimation task.

Ablation Study

Our SUGE method benefits from its three core components. Firstly, it involves neighboring labels combined with pseudo labels and ground-truth labels to assess data uncertainty via triplet-label consistency. Secondly, it designs image and label confidences for label correction and sample weighting. Thirdly, it employs the co-training strategy for alleviating the negative effects of label noise. We conduct ablation experiments using the Gaze360 baseline model on the EyeDiap datasets to reveal the effects of these components.

Neighboring Labeling. To validate the importance of introducing neighboring labels, we conduct two ablations. In the first ablation, we completely remove the neighboring labeling module, and only perform label correction based on the consistency between ground truth labels and pseudo labels, which is similar to the DivideMix (Li, Socher, and Hoi 2020) approach. In the second ablation, we only remove the reconstruction weighting calculation and directly calculate the average of neighbors' labels to obtain the neighboring labels.

In Table 3, we can see that removing the neighboring labeling modules (1st row) leads to an obvious performance drop (0.11°). Besides, the weighting reconstruction strategy (2nd row) can also bring a performance gain (0.05°).

Method	Angle Error (°)
w/o neighboring labeling	5.16
w/o reconstruction weighting	5.10
w/o sample weighting	5.10
w/o label correction	5.17
w/o sample weighting & label correction	5.23
w/o co-training	5.14
w/o entire label compositions	5.10
SUGE	5.05

Table 3: Ablation studies on EyeDiap dataset.

Label correction and sample weighting. To evaluate the individual contributions of the label correction and sample weighting modules, we separately remove sample weighting and label correction modules. Besides, we also exclude both label correction and sample weighting, which can be viewed as an ensemble of two networks. Table 3 shows the results. It can be seen that removing sample weighting (3rd row) results in a performance drop (0.05°), and removing label correction (4th row) brings a larger drop (0.12°). Obviously, combining two strategies for training (5th row) achieves the best performance gain (0.18°), demonstrating the necessity of the label correction and sample weighting modules.

Co-training strategy. To mitigate accumulated errors, we adopt the co-training strategy where two networks feed data samples for each other during training. To validate the effect of co-training strategy, we directly employ the self-training strategy where two networks are trained separately. The results presented in the 6th row of Table 3 indicate that the angle error of self-training is higher than co-training (0.09°), confirming the effectiveness of the co-training strategy.

Label compositions for GT label correction To evaluate the necessity of using the entire label composition in equation 14, we conduct an experiment using its subset, represented as $\hat{Y}_i = \Gamma_i^{\text{label}} \cdot Y_i + (1 - \Gamma_i^{\text{label}}) \cdot \frac{1}{2}(\hat{Y}_i^n + \hat{Y}_i^p)$. Table 3 illustrates that utilizing this subset (7th row) results in a higher angular error (0.05°), confirming the importance of utilizing the entire label composition.

Parameter	Value	Angle Error (°)
K	2	5.08
	4	5.05
	6	5.09
τ^{image}	0.4	5.11
	0.5	5.05
	0.6	5.09
τ^{label}	0.4	5.14
	0.5	5.05
	0.6	5.13

Table 4: Parameter analysis on EyeDiap dataset.

Parameter Sensitivity Analysis

Our SUGE method primarily consists of three hyperparameters: the number of neighbors K , label confidence threshold τ^{label} , and image confidence threshold τ^{image} . We analyze the impact of hyperparameter settings in this experiment. As shown in the Table 4, we can observe that within a certain range, adjusting the hyperparameters has a relatively stable impact on the performance, and $K = 4, \tau^{\text{image}} = 0.5, \tau^{\text{label}} = 0.5$ achieves the best performance.

Visualization Results

In this subsection, we illustrate the effectiveness of our approach in reducing uncertainty of gaze estimation by visualizing some samples, where either their label confidences are set to 0 during training, indicating that these samples are initially annotated as wrong labels, and will be rectified with correct gaze directions, or their image confidences are set to 0, showing that these are low-quality samples and will be discarded during training. As shown in the left of Fig. 4, our method accurately identifies and corrects wrong annotated labels. Moreover, as depicted in the right of Fig. 4, our approach demonstrates a strong capability to identify low-quality images caused by various adverse factors.

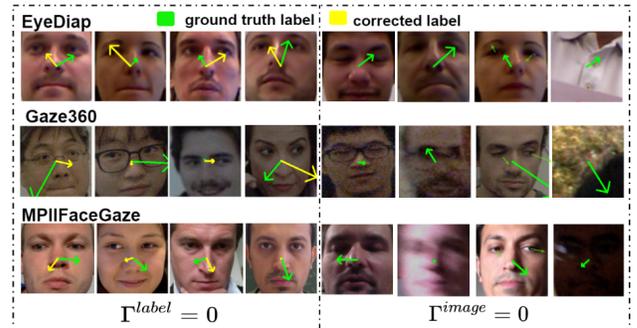


Figure 4: Visualization of samples with label confidences set to 0 on the left figure and image confidences set to 0 on the right figure. These results are sourced from the EyeDiap dataset (folds 0, 1, 3 as the training set), Gaze360 dataset, and MPIIFaceGaze dataset (users 1-14 as the training data), during the initial epoch after warm-up.

Conclusion

In this study, we firstly discover the data uncertainty caused by data collection and annotation process in gaze estimation datasets, which is ignored by existing works. To reduce the negative effects of low-quality images and incorrect labels, we propose a novel approach named SUGE, which adopts the triplet-label consistency to estimate the uncertainty and utilizes it to guide the training process. The comprehensive experiments conducted on popular benchmarks demonstrate that our method significantly improves the performance of gaze estimation.

Acknowledgements

This work is supported by National Natural Science Foundation of China (62271042, 62376021, 62302032, 62106017), Beijing Natural Science Foundation (M22022, L211015, 4232032), and Hebei Natural Science Foundation (F2022105018).

References

- Alberto Funes Mora, K.; and Odobez, J.-M. 2014. Geometric generative gaze estimation (g3e) for remote rgb-d cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1773–1780.
- Cai, X.; Zeng, J.; Shan, S.; and Chen, X. 2023. Source-Free Adaptive Gaze Estimation by Uncertainty Reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22035–22045.
- Cheng, Y.; Huang, S.; Wang, F.; Qian, C.; and Lu, F. 2020. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10623–10630.
- Cheng, Y.; and Lu, F. 2022. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 3341–3347. IEEE.
- Cheng, Y.; Wang, H.; Bao, Y.; and Lu, F. 2021. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*.
- Dias, P. A.; Malafronte, D.; Medeiros, H.; and Odone, F. 2020. Gaze estimation for assisted living environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 290–299.
- Fischer, T.; Chang, H. J.; and Demiris, Y. 2018. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European conference on computer vision (ECCV)*, 334–352.
- Friend, M.; Reising, M.; and Cook, L. 1993. Co-teaching: An overview of the past, a glimpse at the present, and considerations for the future. *Preventing School Failure: Alternative Education for Children and Youth*, 37(4): 6–10.
- Funes Mora, K. A.; Monay, F.; and Odobez, J.-M. 2014. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*, 255–258.
- Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.
- Hansen, D. W.; and Ji, Q. 2009. 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.
- Jiang, M.; and Zhao, Q. 2017. Learning visual attention to identify people with autism spectrum disorder. In *Proceedings of the IEEE international conference on computer vision*, 3267–3276.
- Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; and Torralba, A. 2019. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6912–6921.
- Kim, J.; Stengel, M.; Majercik, A.; De Mello, S.; Dunn, D.; Laine, S.; McGuire, M.; and Luebke, D. 2019. Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–12.
- Li, J.; Socher, R.; and Hoi, S. C. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.
- Li, Y.; Han, H.; Shan, S.; and Chen, X. 2023. DISC: Learning from Noisy Labels via Dynamic Instance-Specific Selection and Correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24070–24079.
- Liu, W.; Li, M.; and Yi, L. 2016. Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research*, 9(8): 888–898.
- Majaranta, P.; and Bulling, A. 2014. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*, 39–65. Springer.
- Nonaka, S.; Nobuhara, S.; and Nishino, K. 2022. Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2192–2201.
- O Oh, J.; Chang, H. J.; and Choi, S.-I. 2022. Self-attention with convolution and deconvolution for efficient eye gaze estimation from a full face image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4992–5000.
- Park, S.; Aksan, E.; Zhang, X.; and Hilliges, O. 2020. Towards end-to-end video-based eye-tracking. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 747–763. Springer.
- Patney, A.; Kim, J.; Salvi, M.; Kaplanyan, A.; Wyman, C.; Benty, N.; Lefohn, A.; and Luebke, D. 2016. Perceptually-based foveated virtual reality. In *ACM SIGGRAPH 2016 emerging technologies*, 1–2.
- Permuter, H.; Francos, J.; and Jermyn, I. 2006. A study of Gaussian mixture models of color and texture features for image classification and segmentation. *Pattern recognition*, 39(4): 695–706.
- Rahal, R.-M.; and Fiedler, S. 2019. Understanding cognitive and affective mechanisms in social psychology through eye-tracking. *Journal of Experimental Social Psychology*, 85: 103842.
- Smith, B.; Yin, Q.; Feiner, S.; and Nayar, S. 2013. Gaze Locking: Passive Eye Contact Detection for Human-Object Interaction. In *ACM Symposium on User Interface Software and Technology (UIST)*, 271–280.

- Sugano, Y.; Matsushita, Y.; and Sato, Y. 2014. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1821–1828.
- Valenti, R.; Sebe, N.; and Gevers, T. 2011. Combining head pose and eye location information for gaze estimation. *IEEE Transactions on Image Processing*, 21(2): 802–815.
- Wang, K.; Peng, X.; Yang, J.; Lu, S.; and Qiao, Y. 2020. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6897–6906.
- Wei, Q.; Feng, L.; Sun, H.; Wang, R.; Guo, C.; and Yin, Y. 2023. Fine-grained classification with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11651–11660.
- Zhang, J.; Fan, D.-P.; Dai, Y.; Anwar, S.; Saleh, F. S.; Zhang, T.; and Barnes, N. 2020a. UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8582–8591.
- Zhang, X.; Park, S.; Beeler, T.; Bradley, D.; Tang, S.; and Hilliges, O. 2020b. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, 365–381. Springer.
- Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017a. It’s written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 51–60.
- Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017b. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1): 162–175.
- Zhou, C.; Huang, Y.; Pu, M.; Guan, Q.; Huang, L.; and Ling, H. 2023. The Treasure Beneath Multiple Annotations: An Uncertainty-aware Edge Detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15507–15517.
- Zhu, Z.; and Ji, Q. 2007. Novel eye gaze tracking techniques under natural head movement. *IEEE Transactions on biomedical engineering*, 54(12): 2246–2260.