PureGaze: Purifying Gaze Feature for Generalizable Gaze Estimation

Yihua Cheng[1], **Yiwei Bao**[1], **Feng Lu**[1, 2*]

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

Abstract

Gaze estimation methods learn eye gaze from facial features. However, among rich information in the facial image, real gaze-relevant features only correspond to subtle changes in eye region, while other gaze-irrelevant features like illumination, personal appearance and even facial expression may affect the learning in an unexpected way. This is a major reason why existing methods show significant performance degradation in cross-domain/dataset evaluation. In this paper, we tackle the cross-domain problem in gaze estimation. Different from common domain adaption methods, we propose a domain generalization method to improve the cross-domain performance without touching target samples. The domain generalization is realized by gaze feature purification. We eliminate gaze-irrelevant factors such as illumination and identity to improve the cross-domain performance. We design a plug-and-play self-adversarial framework for the gaze feature purification. The framework enhances not only our baseline but also existing gaze estimation methods directly and significantly. To the best of our knowledge, we are the first to propose domain generalization methods in gaze estimation. Our method achieves not only state-of-the-art performance among typical gaze estimation methods but also competitive results among domain adaption methods. The code is released in https://github.com/yihuacheng/PureGaze.

Introduction

Human gaze implicates important cues for understanding human cognition (Rahal and Fiedler 2019) and behavior (Dias et al. 2020). It enables researchers to gain insights into many areas such as saliency detection (Wang et al. 2019b; Wang and Shen 2018), virtual reality (Xu et al. 2018) and first-person video analysis (Yu et al. 2020). Recently, appearance-based gaze estimation with deep learning becomes a hot topic. They leverage convolutional neural networks (CNNs) to estimate gaze from human appearance (Park, Spurr, and Hilliges 2018; Xiong, Kim, and Singh 2019), and achieve accurate performance.

CNN-based gaze estimation requires a large number of samples for training. But collecting gaze sample is difficult and time-consuming. This challenge can be ignored in a fixed environment, but becomes a bottleneck when gaze estimation is required in a new environment. The changed environment brings many unexpected factors such as different illumination, thus degrades the performance of pretrained model. Recent methods usually handle the cross-environment problem[1] as a domain adaption problem. Researches aim to adapt the model trained in source domains to target domains (Zhang et al. 2018; Kellnhofer et al. 2019). However, these methods usually require target samples and time-consuming setup. These requirements greatly harm the flexibility of methods.

In this paper, we innovate a new direction to solve the problem. We propose a domain-generalization method for improving the cross-domain performance. Our method does not require any images or labels in target domains, but aims to learn a generalized model in the source domain for any "unseen" target domains. We notice *the intrinsic gaze pattern is similar in all domains, but there are domain differences in gaze-irrelevant factors such as illumination and identity*. These factors are usually domain-specific, and directly blend in captured images. The in-depth fusion makes these factors difficult to be eliminated during feature extrac-

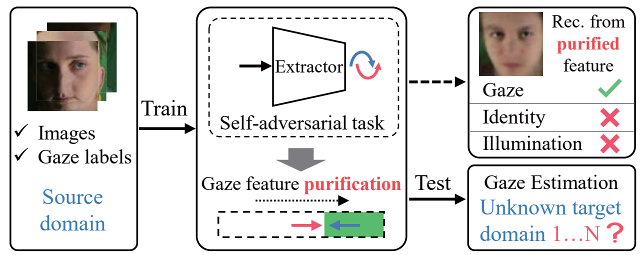


Figure 1: We propose a domain-generalization framework for gaze estimation. Our method is only trained in the source domain and brings improvement in all unknown target domains. The key idea of our method is to purify the gaze feature with self-adversarial framework. The visualization result shows gaze-irrelevant factors such as illumination and identity are eliminated from the extracted feature.

*Corresponding author.
This work was supported by National Natural Science Foundation of China (NSFC) under grant 61972012.

[1]We refer it as cross-domain/dataset problem in the rest.

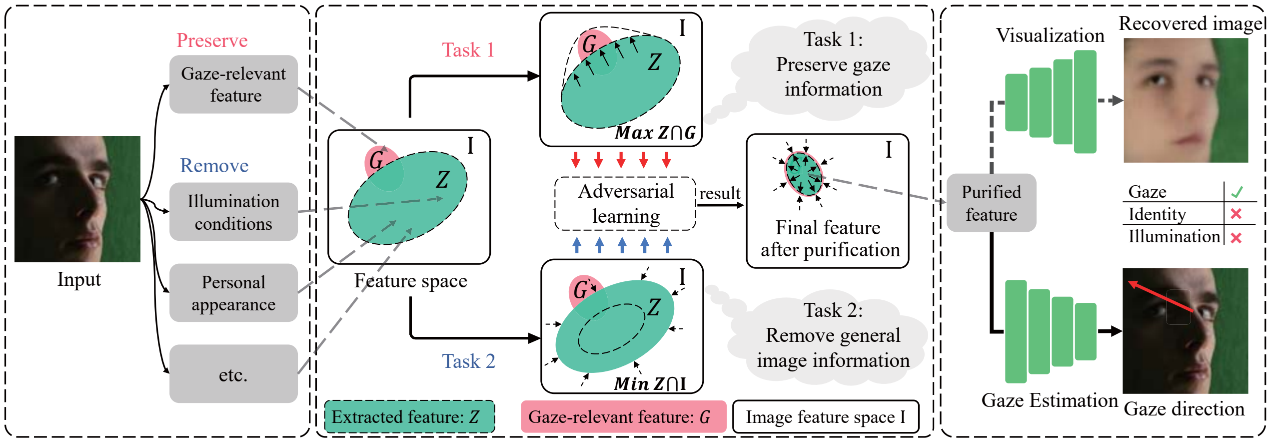


Figure 2: Overview of the gaze feature purification. Our goal is to preserve the gaze-relevant feature and eliminate gaze-irrelevant features. Therefore, we define two tasks, which are to preserve gaze information and to remove general facial image information. The two tasks are not cooperative but adversarial to purify feature. Simultaneously optimizing the two tasks, we implicitly purify the gaze feature without defining gaze-irrelevant feature.

tion. As a result, the trained model usually learns a joint distribution of gaze and these factors, *i.e*., overfit in source domain, and naturally cannot perform well in target domains.

As shown in Fig. 1, the key idea of our method is to purify gaze feature, *i.e*., we eliminate gaze-irrelevant factors such as illumination and identity. The purified feature is more generalized than original feature, and naturally brings improvement in cross-domain performance. To be specific, we propose a plug-and-play self-adversarial framework. As shown in Fig. 2, the framework contains two tasks, which are to preserve gaze information and to remove general facial image information. Simultaneously optimizing the two tasks, we implicitly purify the gaze feature without defining gaze-irrelevant feature. In fact, it is also non-trivial to define all gaze-irrelevant features. We also realize the framework with a practical neural network. As shown in Fig. 3, the two tasks are respectively approximated as a gaze estimation task and an adversarial reconstruction task. We propose the final PureGaze to simultaneously perform the two tasks to purified the gaze feature. The PureGaze contains a plug-and-play SA-Module, which can be used to enhance existing gaze estimation methods directly and significantly.

The contributions of this work are threefold:

- We propose a plug-and-play domain-generalization framework for gaze estimation methods. It improves the cross-dataset performance without knowning the target dataset or touching any new samples. To the best of our knowledge, it is the first domain-generalization framework in gaze estimation.
- The domain-generalizability comes from our proposed gaze feature purification. We design a self-adversarial framework to purify gaze features, which eliminates the gaze-irrelevant factors such as illumination and identity. The purification is easily explainable via visualization as shown in the experiment.
- Our method achieves state-of-the-art performance in many benchmarks. Our plug-and-play module also enhances existing gaze estimation methods significantly.

Related Works

Typical Gaze Estimation. Recently, many gaze estimation methods are proposed. Cheng *et al*. explore the two-eye asymmetry (Cheng, Lu, and Zhang 2018; Cheng et al. 2020b). Park *et al*. generate pictorial gaze representation to handle subject variance (Park, Spurr, and Hilliges 2018). Fisher *et al*. leverage two VGG networks to process two eye images (Fischer, Jin Chang, and Demiris 2018). Zhang *et al*. utilize attention mechanism to weight facial feature (Zhang et al. 2017). Chen *et al*. leverage dilated convolution to estimate gaze (Chen and Shi 2019). Bao *et al*. leverage face and eye images to estimate point of gaze (Bao et al. 2020). Zheng *et al*. propose a gaze/head redirection network and use generated images for data augmentation (Zheng et al. 2020). Cheng *et al*. estimate gaze from facial images and refine the gaze with eye images (Cheng et al. 2020a).

Cross-domain Gaze Estimation. Zhang *et al*. (Zhang et al. 2018) fine tune the pre-trained model in target domain. Wang *et al*. (Wang et al. 2019a) and Kellnhoder *et al*. (Kellnhofer et al. 2019) propose to use adversarial learning to align the features in the source and target domain. Liu *et al*. (Liu et al. 2021) propose an ensemble of networks that learn collaboratively with the guidance of outliers. These methods utilize data from target domain, which is not always user-friendly.

Overview

Definition of the Purification

We first formulate the proposed self-adversarial framework in this section. Without loss of generality, we formulate the gaze estimation problem as

$$\mathbf{g} = F_{\phi}(E_{\theta}(\boldsymbol{I})), \tag{1}$$

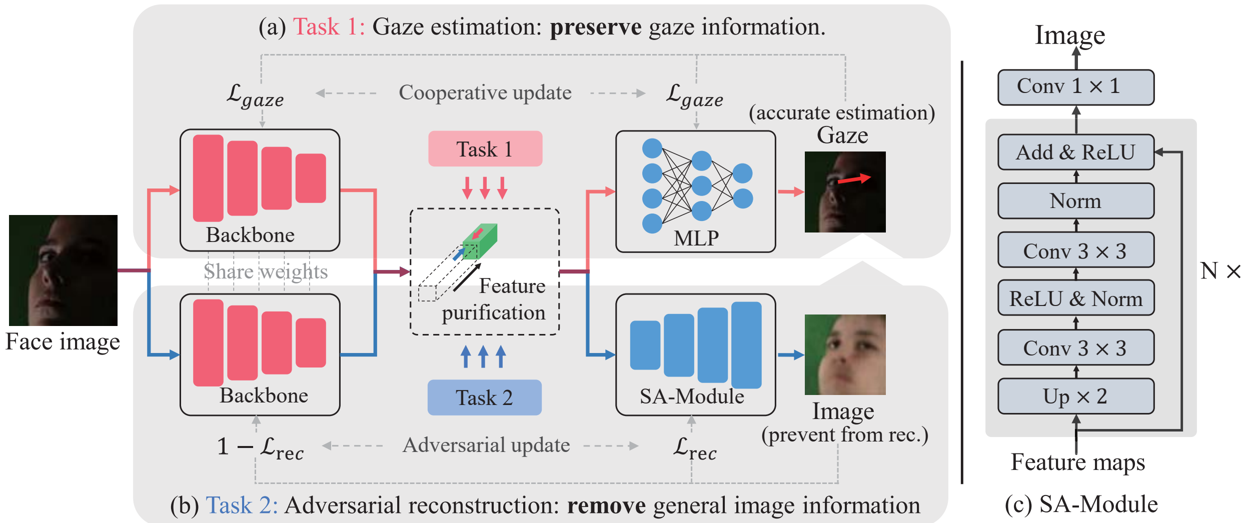


Figure 3: The architecture of PureGaze. It is consisted of two share-weight backbones (ResNet-18) for feature extraction, one two-layer MLP (Muti-layer Perception) for gaze estimation, and one SA-Module (N=5) for recovering images. The backbone and MLP are cooperative to preserve gaze information, *i.e.*, perform gaze estimation, while the backbone and SA-Module are adversarial to remove general image information, *i.e.*, perform adversarial reconstruction. The backbone simultaneously performs the two tasks while the two tasks are not cooperative but adversarial. The backbone performs adversarial learning with itself to purify extracted feature.

where E_θ is a feature extraction function (*e.g.* neural networks), F_ϕ is a regression function, $\boldsymbol{I}$ is a face/eye appearance and $\mathbf{g}$ is a estimated gaze. We use $\boldsymbol{Z}$ to denote the extracted feature, *i.e.*, $\boldsymbol{Z} = E_\theta(\boldsymbol{I})$.

We slightly abuse the notation I to represent a set of all features in one image. We can simply divide the whole image feature into two subsets, gaze-relevant feature G and gaze-irrelevant feature N. It is easy to get the relation:

$$G \cup N = I \quad and \quad G \cap N = \varnothing \tag{2}$$

Our goal is to find the optimal E_{θ^*} to extract purified feature Z^*, where Z^* does not contain gaze-irrelevant feature, *i.e.*, $Z^* \cap N = \varnothing$. Besides, we believe the feature which has weak relations with gaze should be also eliminated to improve generalization.

Self-Adversarial Framework

As shown in Fig. 2, we design two tasks for feature purification. The first task is to minimize the mutual information (MI) between image feature and the extracted feature, *i.e.*,

$$\theta^* = \arg\min_\theta H(I, Z) \tag{3}$$

The function $H(X, Y)$ computes the MI between X and Y. It indicates the relation between X and Y, *e.g.* , $H(X, Y) = 0$ if X is independent with Y. This task also means the extracted feature should contain less image information.

The other task is to maximize the MI between gaze-relevant feature and extracted feature, *i.e.*,

$$\theta^* = \arg\max_\theta H(G, Z), \tag{4}$$

This constraint means the extracted feature should contain more gaze-relevant information.

Learning to Purify in the Framework

We simultaneously solve Equ. (3) and Equ. (4). In other words, the extracted feature needs to contain more gaze information (Equ. (4)) and less image information (Equ. (3)). The two optimization tasks compose a self-adversarial framework on the extracted feature. During the optimization, gaze-irrelevant feature is eliminated to satisfy Equ. (3) and gaze-relevant feature is preserved to satisfy Equ. (4). In the other word, we purify extracted feature with the self-adversarial framework.

In addition, Equ. (3) and Equ. (4) implicate the minimax problem of $H(G, Z)$. It is intuitive that the extracted feature will gradually discard some gaze-relevant information to decrease image information, *i.e.*, to satisfy Equ. (3). Meanwhile, to satisfy Equ. (4), the feature having weak relations with gaze will be discarded first.

PureGaze

In the previous section, we propose two key tasks, *i.e.*, Equ. (3) and Equ. (4). The two tasks compose a self-adversarial framework to purify feature. In this section, we propose PureGaze based on the framework. We realize the two tasks with two practical tasks, gaze estimation and adversarial reconstruction. We also propose two loss function for the framework.

Gaze estimation: We use gaze estimation tasks to preserve gaze information in the extracted feature, *i.e.*, Equ. (4).

In fact, the task can be realized with any gaze estimation network. We simply divide gaze estimation networks into two subnets, backbone for extracting feature and MLP for regressing gaze from the feature (Fig. 3(a)). We use a gaze loss function $\mathcal{L}_{gaze}$ such as L1 loss to optimize the two subnets. The two subnets cooperate to preserve gaze information.

Adversarial reconstruction: We propose an adversarial reconstruction task to remove general image information from extracted feature, *i.e.*, Equ. (3). Our assumption is that if the reconstruction network cannot recover input images from extracted feature, it means the extracted feature contains no image information.

Therefore, we first propose a SA-Module for reconstruction as shown in Fig. 3(c). It contains a block for upsampling and a 1×1 convolution layer to align channels. Further, the network architecture for adversarial reconstruction is shown in Fig. 3(b). We use a backbone for feature extraction and SA-Module for recovering images. We assign adversarial losses to the backbone and SA-Module. The SA-Module tires to reconstruct images and is optimized with an reconstruction loss $\mathcal{L}_{rec}$ such as pixel-wise MSE Loss. The backbone tires to prevent the reconstruction. We use an adversarial loss $\mathcal{L}_{adv}$ to optimize it, where

$$\mathcal{L}_{adv} = 1 - \mathcal{L}_{rec}. \tag{5}$$

It is obvious that the backbone and the SA-Module are adversarial in reconstruction, *i.e.*, the backbone finally removes general image information from extracted feature.

Architecture of PureGaze

The architecture of PureGaze is shown in the left part of Fig. 3. We respectively build two networks for gaze estimation and adversarial reconstruction with the same backbone, and share the weight of two backbones.

In general, PureGaze contains three networks, which are a backbone for feature extraction, a MLP for gaze estimation and a SA-Module for image reconstruction. The loss functions of the three parts are

$$\mathcal{L}_{SA} = \mathcal{L}_{rec}. \tag{6}$$

$$\mathcal{L}_{MLP} = \mathcal{L}_{gaze}. \tag{7}$$

$$\mathcal{L}_{backbone} = \alpha\mathcal{L}_{adv} + \beta\mathcal{L}_{gaze}. \tag{8}$$

where α and β are hyper-parameters. In this paper, we use L1 Loss for gaze estimation and pixel-wise MSE for reconstruction:

$$\mathcal{L}_{gaze} = \left\|\mathbf{g} - \hat{\mathbf{g}}\right\|_1. \tag{9}$$

$$\mathcal{L}_{rec} = \left\|I - \hat{I}\right\|_2. \tag{10}$$

Purifying Feature in Training: PureGaze uses one backbone to extract feature. The backbone has two goals, minimizing $\mathcal{L}_{gaze}$ and minimizing $\mathcal{L}_{adv}$. Minimizing $\mathcal{L}_{gaze}$ means the backbone should extract gaze-irrelevant feature, while minimizing $\mathcal{L}_{adv}$ means the backbone should not extract any image feature. The two goals are not cooperative but adversarial, and compose an adversarial learning to purify the extract feature. In addition, $\mathcal{L}_{adv}$ is easily satisfied with learning a local optimal solution to cheat the SA-Module. We design another task $\mathcal{L}_{rec}$ to against $\mathcal{L}_{adv}$ to avoid the local optimal solution. The two novel adversarial tasks both are important parts in PureGaze.

Local Purification Loss

It is intuitive that eye region is more important than other facial regions for gaze estimation. Therefore, we want PureGaze to pay more attention to purify the feature of eye region. One simple solution is to directly use eye images as input. However, we believe it is unreasonable since other facial regions also provide useful information.

As a result, we propose the local purification loss (LP-Loss). We use an attention map to focus the purification on a local region. Note that, the attention map is only applied to $\mathcal{L}_{adv}$, *i.e.*, the loss function of the backbone is modified as

$$\mathcal{L}_{backbone} = \alpha\mathbb{E}[\boldsymbol{M} * (\mathbf{1} - (I_i - \hat{I}_i)^2)] + \beta\mathcal{L}_{gaze}. \tag{11}$$

where $\boldsymbol{M}$ is the attention map, and $*$ means element-wise multiplication. In this paper, we use mixed guassian distribution to generate the attention map. We use the coordinates of two eye centers as mean values, and the variance $\Sigma = diag(\sigma^2, \sigma^2)$ of the distribution can be customized.

The loss function can be understood as following. On the one hand, LP-Loss focuses the purification on eye region. On the other hand, LP-Loss does not change the loss function of gaze estimation, *i.e.*, $\mathcal{L}_{gaze}$. Human gaze is estimated from whole face images rather than weighted face images.

Truncated Adversarial Loss

The adversarial reconstruction task plays an important role in the PureGaze. It ensures the extracted feature contains less image feature. In PureGaze, we minimize $\mathcal{L}_{adv}$ to prevent the reconstruction. A smaller value of $\mathcal{L}_{adv}$ indicates a larger pixel difference between the generated and original images. However, we think it is redundant to produce a very large pixel difference. The reason is that $\mathcal{L}_{adv}$ is designed to prevent the reconstruction rather than recover an "inverse" version of the original image.

Therefore, we further propose the truncated adversarial loss (TA-Loss). We use a threshold k to truncate the adversarial loss $\mathcal{L}_{adv}$. In one word, $\mathcal{L}_{adv}$ will be zero if the pixel difference is larger than k. The final loss function of the backbone is:

$$\mathcal{L}_{backbone} = \alpha\mathbb{E}[\boldsymbol{M} * \mathbb{1}_{[\mathbf{1} - \left\|I_i - \hat{I}_i\right\|_2 > k]} * (\mathbf{1} - (I_i - \hat{I}_i)^2)] + \beta\mathcal{L}_{gaze}. \tag{12}$$

where $\mathbb{1}$ is the indicator function and k is the threshold.

Experiments

Data-preprocessing

Task definitions: We use Gaze360 (Kellnhofer et al. 2019) and ETH-XGaze (Zhang et al. 2020) as training set, since they have a large number of subjects, various gaze range and head pose. We test our model in two popular datasets, which are MPIIGaze (Zhang et al. 2017) and EyeDiap (Funes Mora, Monay, and Odobez 2014). We totally conduct four cross-dataset tasks, and denote them as E (ETH-XGaze)→M (MPIIGaze), E→D (EyeDiap), G (Gaze360)→M, G→D.

Data Preparing. We follow (Cheng et al. 2021) to prepare datasets. Gaze360 (Kellnhofer et al. 2019) dataset contains a total of 172K images from 238 subjects. Note that some

Category	Methods	Target Samples	G→M	G→D	E→M	E→D
without adaption	RT-Gene (Fischer, Jin Chang, and Demiris 2018)	*N/A*	21.81°	38.60°	-	-
	Dilated-Net (Chen and Shi 2019)	*N/A*	18.45°	23.88°	-	-
	Full-Face (Zhang et al. 2017)	*N/A*	11.13°	14.42°	12.35°	30.15°
	CA-Net (Cheng et al. 2020a)	*N/A*	27.13°	31.41°	-	-
	ADL* (Kellnhofer et al. 2019)	*N/A*	11.36°	11.86°	7.23°	8.02°
	Baseline (ours)	*N/A*	9.89°	11.42°	8.13°	7.74°
	PureGaze (ours)	*N/A*	**9.28°**	**9.32°**	**7.08°**	**7.48°**
with adaption	ADL (Kellnhofer et al. 2019)	> 100	9.70°	10.28°	5.48°	16.11°
	DAGEN (Guo et al. 2020)	~ 500	6.61°	12.90°	6.16°	9.73°
	ADDA (Tzeng et al. 2017)	~ 500	8.76°	14.80°	6.33°	7.90°
	GVBGD (Cui et al. 2020)	~ 1000	7.64°	12.44°	6.68°	7.27°
	UMA (Cai, Lu, and Sato 2020)	~ 100	8.51°	19.32°	7.52°	12.37°
	PNP-GA (Liu et al. 2021)	< 100	6.18°	7.92°	5.53°	**5.87°**
	Fine-tuned Baseline	< 100	5.28°	7.66°	5.68°	7.26°
	Fine-tuned PureGaze	< 100	**5.20°**	**7.36°**	**5.30°**	6.42°

Table 1: Performance comparison with SOTA methods. PureGaze shows best performance among typical gaze estimation methods (*w/o* adaption), and has competitive result among domain adaption methods. Note that, PureGaze learns one optimal model for four tasks, while domain adaption methods need to learn a total of four models. This is an advantage of our method.

of the images in Gaze360 only captured the back side of the subject. These images is not suitable for appearance-based methods. Therefore, we first clean the dataset with a simple rule. We remove the images without face detection results based on the provided face detection annotation. ETH-XGaze (Zhang et al. 2020) contains a total of 1.1M images from 110 subjects. It provides a training set containing 80 subjects. We split 5 subjects for validation and others are used for training. MPIIGaze (Zhang et al. 2017) is prepared based on the standard protocol. We collect a total of 45K images from 15 subjects. EyeDiap (Funes Mora, Monay, and Odobez 2014) provide a total of 94 video clips from 16 subjects. We follow the common steps to prepare the data as in (Zhang et al. 2017; Cheng et al. 2020a). Concretely, we select the VGA videos of screen targets session and sample one image every fifteen frames. We also truncate the data to ensure the number of images from each subject is the same.

Data rectification. Data rectification is performed to simplify the gaze estimation task. We follow (Sugano, Matsushita, and Sato 2014) to process MPIIGaze and (Zhang, Sugano, and Bulling 2018) to process EyeDiap. ETH-XGaze is already rectified before publication. Gaze360 rectifies their gaze directions to cancel the effect caused by camera pose. We directly use their provided data.

Comparison Methods

Baseline: We remove the SA-Module in PureGaze. The new network is denoted as Baseline. It is obvious the performance difference between PureGaze and Baseline is caused by SA-Module. We also denote the feature extracted by Baseline as original feature, and the feature extracted by PureGaze as purified feature.

Gaze estimation methods: We compare our method with four methods, which are Full-Face (Zhang et al. 2017), RT-Gene (Fischer, Jin Chang, and Demiris 2018), Dilated-Net (Chen and Shi 2019) and CA-Net (Cheng et al. 2020a). These methods all perform well in within-dataset evaluation. We implement Full-Face and Dilated-Net using Pytorch, and use the official code of the other two methods.

Domain adaption methods: We also compare our method with domain adaption methods for reference. In fact, it is unfair to compare our method with domain adaption methods since these methods require target samples. Adversarial learning (ADL) (Kellnhofer et al. 2019) is proved useful in gaze estimation, and has a similar feature with our method. We implement ADL for main comparison. We also modify the method as ADL*, where we only use a discriminator to distinguish personal feature in source domains. ADL* does not require target samples as our methods. In addition, we directly report the performance of other domain adaption methods from (Liu et al. 2021) for reference.

Performance Comparison with SOTA Methods

We first conduct experiments in four cross-dataset tasks. The result is shown in Tab. 1. Note that, Dilated-Net, CA-Net and RT-Gene are not applicable in ETH-XGaze, since ETH-XGaze cannot always provide reliable eye images. In addition, ETH-XGaze dataset uses an off-the-shelf ResNet50 as

Methods	G→M	G→D	E→M	E→D
Full-Face	11.13°	14.42	12.35°	30.15°
Full-Face+SA (ours)	9.16°	14.20	11.50°	21.01°
CA-Net	27.13°	31.41°	-	-
CA-Net+SA (ours)	**9.03°**	9.71°	-	-
Baseline (ours)	9.89°	11.42°	8.13°	7.74°
PureGaze (ours)	9.28°	**9.32°**	**7.08°**	**7.48°**

Table 2: We apply the self-adversarial framework into other advanced gaze estimation methods. Our framework directly enhances existing gaze estimation methods. The experiment also provides a more fair comparison with these methods.

baseline. We follow the protocol and replace the backbone in our method and ADL with ResNet50 in ETH-XGaze.

Comparison with typical gaze estimation methods: The second row of Tab. 1 shows the comparison between our method and gaze estimation methods. Our method and compared methods are all trained on source domains and evaluated on target domains. It is obvious current gaze estimation methods usually have bad performance in cross-dataset evaluation. This is because these methods are easily over-fitted in source domains. In contrast, our Baseline has good performance in all tasks due to simple architecture. PureGaze further improves the performance of Baseline and achieves the state-of-the-art performance in all tasks.

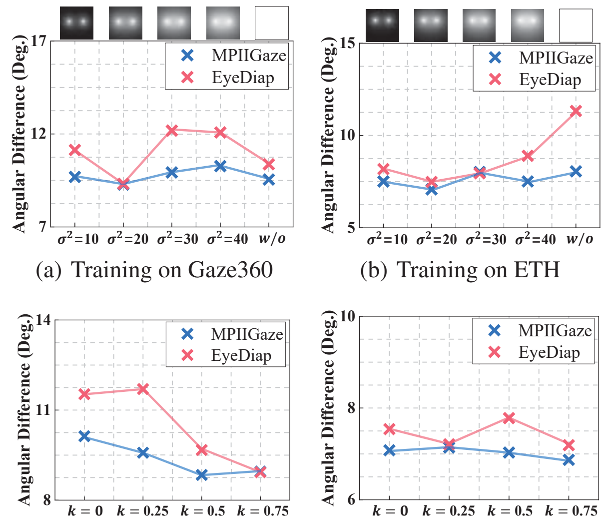
Comparison with domain adaption methods: The third row of Table 1 shows the performance of domain adaption methods. ADL has the same backbone as PureGaze. It improves the performance in three tasks and fails in $E \rightarrow D$. It also has the best performance among compared methods in $E \rightarrow M$. Compared with ADL, PureGaze surpasses the performance of ADL in three tasks without target samples. This proves the effectiveness of our method.

On the other hand, we evaluate the performance of ADL* and show the result in the second row of Table 1. Without target samples, ADL* cannot always bring performance improvement. This is because it is hard to improve performance in all unknown domains, and also demonstrate the value of our method. PureGaze surpasses ADL * in all tasks.

Table 1 also shows the performance of other domain adaption methods. PureGaze shows competitive result among these domain adaption methods without domain adaption. In addition, we also provide a simple application of PureGaze, where we sample 5 images per person from target domains to fine tune PureGaze. Fine-tuned PureGaze further improves the performance with a fast calibration/adaption. More valuable, we observe fine-tuned PureGaze also has better performance than fine-tuned Baseline. This proves PureGaze learns a better feature representation.

Plug Existing Gaze Estimation Methods

We also apply our self-adversarial framework into Full-Face (Zhang et al. 2017) and CA-Net (Cheng et al. 2020a). We input their final facial feature maps into SA-Module, and simply add two loss functions, $\mathcal{L}_{rec}$ and $\mathcal{L}_{adv}$.



(a) Training on Gaze360 (b) Training on ETH

(c) Training on Gaze360 (d) Training on ETH

Figure 4: We evaluate the hyper-parameters of the two loss function. The first row shows the result about LP-Loss, where w/o means we ablate the loss. The second row shows the result about TA-Loss, where $k = 0$ means we ablate TA-Loss. Our method is the best when σ^2 is 20 and k is 0.75.

The result is shown in Tab. 2. It is surprising that CA-Net has the worst performance in G→M , while CA-Net+SA has the best performance in G→M. Besides, it also improved by nearly 70% in G→D. Full-Face+SA also shows better performance than Full-Face in all tasks. The experiment provides a more fair comparison with typical gaze estimation methods, and proves the plug-and-play attribute of our self-adversarial framework. Note that, our framework does not require additional inference parameters and training images. This is a key advantage of our method.

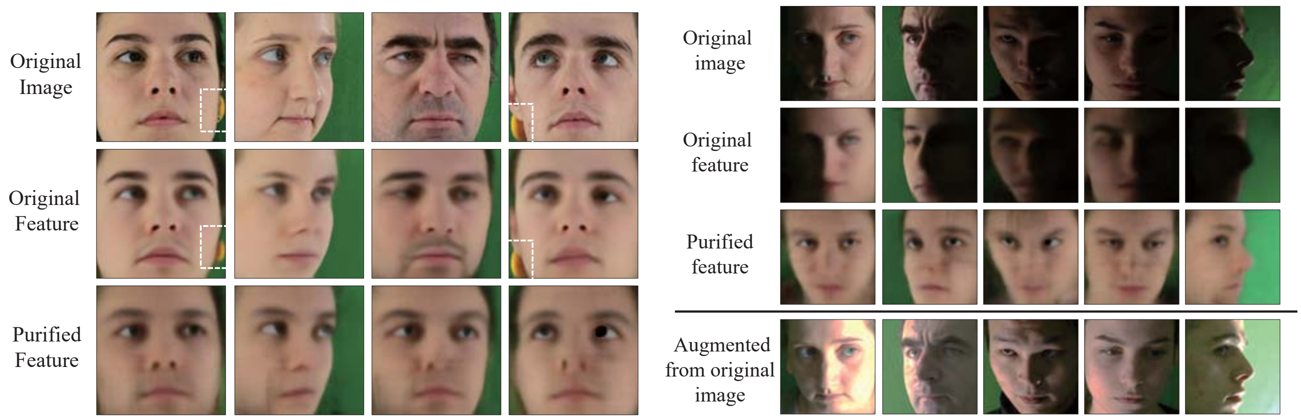
Ablation Study of Two Loss Function

Self-adversarial framework provides a primary architecture of PureGaze while it is rough. We also propose two loss functions (LP-Loss and TA-Loss) to enhance PureGaze. The two loss functions both have hyper-parameters, *i.e.* , variance σ^2 in LP-Loss and threshold k in TA-Loss. We conduct experiments about different hyper-parameters in this section.

As shown in Fig. 4(a) and Fig. 4(b), we set four values for σ^2, which are 10, 20, 30 and 40, and also evaluate the performance without the loss. We illustrate the generated attention maps in the top of the two figures. As for TA-Loss, we set four values for k, which are 0, 0.25, 0.5 and 0.75. $k = 0$ also means we ablate TA-Loss. The results are shown in Fig. 4(c) and Fig. 4(d). It is obvious that the two loss functions both brings performance improvement. When $k = 0.75$ and $\sigma^2 = 20$, PureGaze has best performance.

Visualize Extracted Feature via Reconstruction

To verify the key idea of gaze feature purification, we visualize purified features for further understanding. We provide



(a) Reconstruction by purified feature: less identity. (b) Reconstruction by purified feature: less affected by illumination.

Figure 5: We visualize the purified feature and the original feature via reconstruction. The result clearly explains the purification. Our purified feature contains less gaze-irrelevant feature and naturally improves the cross-domain performance. a) The purified feature contains less identity information than the original feature. Besides, it is interesting that a head rest is captured in the first and fourth columns. The original feature also contains the head rest information while our method eliminates it. b) The purified feature contains less illumination information. We also manually augment each original image in the fourth row. The result shows our method accurately captures the gaze information under the dash area.

reconstruction results of purified features and original features for comparison. We directly show the output of SA-Module to visualize the purified feature. We freeze the parameters of the pre-trained model and simply train a SA-Module to reconstruct images from the original feature.

According to the visualization result shown in Fig. 5, we could easily draw following conclusions:

- The purified feature contains less identity information than original feature. The reconstructed face appearances are approximately the same for each subject.
- The purified feature contains less illumination factors. Besides, it is interesting that our method also recover a bright gaze region accurately from low-light images. This means our method is able to effectively extract gaze information under the dash area.
- Except illumination and identity factors, our method also eliminates other gaze-irrelevant features like the head rest in Fig. 5(a).

Note that our method does not specify eliminated factors. PureGaze automatically purifies the learned feature. This is an advantage of our method since it is non-trivial to manually list all gaze-irrelevant feature.

Discussion

1) Domain generalization. Gaze estimation methods usually have large performance drop when tested in new environment. This feature limits the application of gaze estimation. In this paper, we innovate a new direction to solve the cross-domain problem. Compared with domain adaption (DA) methods, domain generalization (DG) methods are more flexible, *e.g.*, the setup of DA methods usually is time-consuming while DG methods can be directly applied to new domains. But as a trade-off, DG methods usually perform worse than DA methods due to the lack of target domains information. The trade-off between flexibility and accuracy should be considered by researchers.

2) Self-adversarial framework. We propose a self-adversarial framework to learn purified feature. The purified feature improves the cross-domain performance without touching target samples. In fact, our framework can also be considered as a zero-shot cross-domain method since we require no samples in target domains. The zero-shot mechanism is designed based on the observation, gaze pattern is similar in all domains, while some gaze-irrelevant factors are usually domain-specific and bring performance drop. Our method eliminates some gaze-irrelevant feature and naturally improves the cross-domain performance. However, the same as DA methods, our method is unstable in source domains. Our method slightly changes the performance in source domain ($\pm 0.2°$). This is because an over-fitting model can achieve better performance in source domain compared with PureGaze. Learning more generalized model is a future direction of our framework.

Conclusion

In this paper, we innovate a new direction to solve the cross-dataset problem in gaze estimation. We propose a plug-and-play domain-generalization framework. The framework purifies gaze feature to improve the performance in unknown target domains without touching the domain. Experiments show our method achieves state-of-the-art performance among typical gaze estimation methods and also has competitive result compared with domain adaption methods.

References

Bao, Y.; Cheng, Y.; Liu, Y.; and Lu, F. 2020. Adaptive Feature Fusion Network for Gaze Tracking in Mobile Tablets. In *International Conference on Pattern Recognition (ICPR)*.

Cai, M.; Lu, F.; and Sato, Y. 2020. Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 14392–14401.

Chen, Z.; and Shi, B. E. 2019. Appearance-Based Gaze Estimation Using Dilated-Convolutions. In Jawahar, C.; Li, H.; Mori, G.; and Schindler, K., eds., *The Asian Conference on Computer Vision*, 309–324. ISBN 978-3-030-20876-9.

Cheng, Y.; Huang, S.; Wang, F.; Qian, C.; and Lu, F. 2020a. A Coarse-to-fine Adaptive Network for Appearance-based Gaze Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Cheng, Y.; Lu, F.; and Zhang, X. 2018. Appearance-Based Gaze Estimation via Evaluation-Guided Asymmetric Regression. In *The European Conference on Computer Vision*.

Cheng, Y.; Wang, H.; Bao, Y.; and Lu, F. 2021. Appearance-based Gaze Estimation With Deep Learning: A Review and Benchmark. *arXiv preprint arXiv:2104.12668*.

Cheng, Y.; Zhang, X.; Lu, F.; and Sato, Y. 2020b. Gaze Estimation by Exploring Two-Eye Asymmetry. *IEEE Transactions on Image Processing*, 29: 5259–5272.

Cui, S.; Wang, S.; Zhuo, J.; Su, C.; Huang, Q.; and Tian, Q. 2020. Gradually vanishing bridge for adversarial domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 12455–12464.

Dias, P. A.; Malafronte, D.; Medeiros, H.; and Odone, F. 2020. Gaze Estimation for Assisted Living Environments. In *The IEEE Winter Conference on Applications of Computer Vision*.

Fischer, T.; Jin Chang, H.; and Demiris, Y. 2018. RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments. In *The European Conference on Computer Vision*.

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 ACM Symposium on Eye Tracking Research & Applications*.

Guo, Z.; Yuan, Z.; Zhang, C.; Chi, W.; Ling, Y.; and Zhang, S. 2020. Domain Adaptation Gaze Estimation by Embedding with Prediction Consistency. In *The Asian Conference on Computer Vision*.

Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; and Torralba, A. 2019. Gaze360: Physically Unconstrained Gaze Estimation in the Wild. In *The IEEE International Conference on Computer Vision*.

Liu, Y.; Liu, R.; Wang, H.; and Lu, F. 2021. Generalizing Gaze Estimation with Outlier-guided Collaborative Adaptation. In *The IEEE International Conference on Computer Vision*.

Park, S.; Spurr, A.; and Hilliges, O. 2018. Deep Pictorial Gaze Estimation. In *The European Conference on Computer Vision*.

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.

Sugano, Y.; Matsushita, Y.; and Sato, Y. 2014. Learning-by-Synthesis for Appearance-based 3D Gaze Estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition*, 7167–7176.

Wang, K.; Zhao, R.; Su, H.; and Ji, Q. 2019a. Generalizing Eye Tracking With Bayesian Adversarial Learning. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, W.; and Shen, J. 2018. Deep Visual Attention Prediction. *IEEE Transactions on Image Processing*, 27(5): 2368–2378.

Wang, W.; Shen, J.; Dong, X.; Borji, A.; and Yang, R. 2019b. Inferring Salient Objects from Human Fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xiong, Y.; Kim, H. J.; and Singh, V. 2019. Mixed Effects Neural Networks (MeNets) With Applications to Gaze Estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Xu, Y.; Dong, Y.; Wu, J.; Sun, Z.; Shi, Z.; Yu, J.; and Gao, S. 2018. Gaze Prediction in Dynamic 360° Immersive Videos. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Yu, H.; Cai, M.; Liu, Y.; and Lu, F. 2020. First- And Third-person Video Co-analysis By Learning Spatial-temporal Joint Attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Zhang, X.; Huang, M. X.; Sugano, Y.; and Bulling, A. 2018. Training Person-Specific Gaze Estimators from User Interactions with Multiple Devices. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ISBN 9781450356206.

Zhang, X.; Park, S.; Beeler, T.; Bradley, D.; Tang, S.; and Hilliges, O. 2020. ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation. In *The European Conference on Computer Vision*.

Zhang, X.; Sugano, Y.; and Bulling, A. 2018. Revisiting Data Normalization for Appearance-Based Gaze Estimation. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*. ISBN 9781450357067.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017. Its written all over your face: Full-face appearance-based gaze estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2299–2308.

Zheng, Y.; Park, S.; Zhang, X.; De Mello, S.; and Hilliges, O. 2020. Self-Learning Transformations for Improving Gaze and Head Redirection. *Advances in Neural Information Processing Systems*.