Test-Time Personalization with Meta Prompt for Gaze Estimation

Huan Liu[1], Julia Qi[1,2*†], Zhenhao Li[1*], Mohammad Hassanpour[1], Yang Wang[3], Konstantinos Plataniotis[4], Yuanhao Yu[1]

[1] Noah's Ark Lab, Huawei, Canada
[2] University of Waterloo, Canada
[3] Department of Computer Science and Software Engineering, Concordia University, Canada
[4] The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Canada
{huan.liu127, zhenhao.li1, mohammad.hassanpour, yuanhao.yu}@huawei.com, j6qi@waterloo.ca, yang.wang@concordia.ca, kostas@ece.utoronto.ca

Abstract

Despite the recent remarkable achievement in gaze estimation, efficient and accurate personalization of gaze estimation without labels is a practical problem but rarely touched on in the literature. To achieve efficient personalization, we take inspiration from the recent advances in Natural Language Processing (NLP) by updating a negligible number of parameters, "prompts", at the test time. Specifically, the prompt is additionally attached without perturbing original network and can contain less than 1% of a ResNet-18's parameters. Our experiments show high efficiency of the prompt tuning approach. The proposed one can be 10 times faster in terms of adaptation speed than the methods compared. However, it is non-trivial to update the prompt for personalized gaze estimation without labels. At the test time, it is essential to ensure that the minimizing of particular unsupervised loss leads to the goals of minimizing gaze estimation error. To address this difficulty, we propose to meta-learn the prompt to ensure that its updates align with the goal. Our experiments show that the meta-learned prompt can be effectively adapted even with a simple symmetry loss. In addition, we experiment on four cross-dataset validations to show the remarkable advantages of the proposed method.

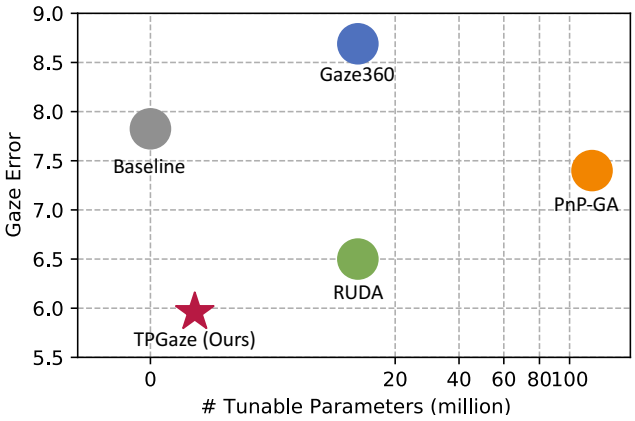

Figure 1: Illustration of performance comparison in terms of gaze error and tunable parameters. Our method achieved the lowest gaze error with negligible tunable parameters required for personalization. Gaze errors are calculated by averaging four cross-dataset validations.

1 Introduction

Gaze, which refers to the direction of an individual's visual focus, is a crucial indicator of human attention. Gaze estimation is a rapidly evolving area of research, with the aim of determining the direction of an individual's gaze based on the positioning and orientation of their eyes and head. Due to its potential applications in diverse fields such as healthcare (Castner et al. 2020), gaming (Burova et al. 2020), and human-computer interaction (Admoni and Scassellati 2017), it has attracted significant attention.

Recent years have witnessed tremendous success of utilizing deep learning in addressing the gaze estimation problem. However, most of deep learning based methods (Krafka et al. 2016; Fischer, Chang, and Demiris 2018; Kellnhofer et al. 2019; Funes Mora, Monay, and Odobez 2014; Zhang et al. 2020, 2017b) essentially learn a mapping on the training data in a supervised manner. Although achieving high accuracy on the training dataset, these methods suffer from performance degradation when tested in real-world scenarios with distribution shift. In addition, collecting labeled data in the real world is extremely difficult, making it challenging to fine-tune. These issues raise concerns about the practical value of purely supervised methods. To address the above concern, recent attempts have focused on gaze estimation in the unsupervised domain adaptation (UDA) setting (Kellnhofer et al. 2019; Wang et al. 2019; Liu et al. 2021b). However, these UDA methods assume the availability of source data, which may not be available in real-world applications due to privacy concerns. In the literature, only few methods, such as (Wang et al. 2022) and (Bao et al. 2022), consider addressing UDA without source data.

In this paper, we study a new variant of UDA, i.e., test-time personalization. With the widespread use of portable

*These authors contributed equally.
†Work done during an internship at Huawei Noah's Ark Lab.

devices such as smartphones, laptops and tablets, personalized experiences have become increasingly necessary. Personalizing gaze estimation based on the user's unique characteristics is essential for enhancing their experience. Unfortunately, current personalized gaze estimation methods usually require either calibration (Krafka et al. 2016) or ground truth gaze labels (Park et al. 2019; Liu et al. 2019; Chen and Shi 2020; Yu, Liu, and Odobez 2019; Ghosh et al. 2022), which may significantly limit their practical deployment. One possible solution is to directly employ (Wang et al. 2022) and (Bao et al. 2022) for the adaptation on personal data. However, applying these methods under test-time personalization settings is non-trivial. They require all parameters to be trainable during the adaptation phase, which may not be practical to be conducted on edge devices at the test time due to computational constraints.

To this end, we propose a **T**est-time **P**ersonalized **Gaze** estimation (**TPGaze**) method by considering both adaptation efficiency and effectiveness. Specifically, we take inspiration from natural language processing research (Li and Liang 2021; Liu et al. 2023b; Lester, Al-Rfou, and Constant 2021) and propose to update a small group of parameters solely, namely the "prompt", while freezing the backbone of the network during personalization. The prompt is person-specific and memory-saving, with a cost of less than 1% of a ResNet-18 model. Ideally, it might be feasible to update the prompt for performance improvements using unsupervised gaze-relevant losses, such as the rotation consistency loss (Bao et al. 2022). These loss functions are carefully yet intuitively designed to be correlated to gaze.

The effectiveness of the losses is primarily demonstrated through empirical studies, such as experimental results. However, it is somewhat hard to guarantee that the gradients with respect to such losses align with the direction that minimizes gaze estimation error for any particular person. To bridge unsupervised losses and gaze error, we propose a meta-learning-based approach that can explicitly associate the two objectives. The goal of meta-learning here is to learn an ideal initialization of the prompt across individuals so that its updates towards lower unsupervised losses are equivalent to updates towards lower gaze estimation error. We will show that meta-learning is very effective even with a simple left-right symmetry loss (Kellnhofer et al. 2019).

Our main contributions are summarized as follows: 1) We propose an efficient method for test-time personalized gaze estimation that achieves fast adaptation by leveraging prompt. 2) Our proposed method employs meta-learning to initialize the prompt, explicitly ensuring that test-time prompt updates result in reduced gaze estimation error. 3) Extensive experimental results demonstrate the effectiveness of our method. To illustrate, Figure 1 presents a preview of our superior performance.

2 Related Work

2.1 Appearance-based Gaze Estimation

Research on gaze-direction-from-eye-appearance has been ongoing for over a century (Wollaston 1824). Early approaches (Wood and Bulling 2014; Reale, Hung, and Yin 2010) mainly rely on constructing geometric eye models from images. In the recent decade, deep learning on labeled datasets (Zhang et al. 2017b; Krafka et al. 2016; Zhang et al. 2017a; Funes Mora, Monay, and Odobez 2014; Kellnhofer et al. 2019; Zhang et al. 2020) has been a game changer to the field, eliminating the need to explicitly construct eye models. Deep neural network (DNN) models present a solid ability to learn rich gaze-relevant features from supervised learning, achieving breakthroughs in estimation accuracy (Krafka et al. 2016; Cheng, Lu, and Zhang 2018; Park, Spurr, and Hilliges 2018; Cheng et al. 2020b,a; Biswas et al. 2021; Lian et al. 2018, 2019a). Besides, GazeNeRF (Ruzzi et al. 2023) propose a 3D-aware design by incorporating neural radiance fields to synthesize more samples for effective supervised training.

Transfer learning techniques, such as domain adaptation, have been adopted in recent gaze estimation approaches to improve the performance of a DNN model across datasets. Most follow the UDA setting (Kellnhofer et al. 2019; Wang et al. 2019; Liu et al. 2021b; Bao et al. 2022; Guo et al. 2020), in which labeled source domain data and unlabeled target domain data are accessible during adaptation. CRGA (Wang et al. 2022), RUDA (Bao et al. 2022) and UnReGA(Cai et al. 2023) perform source-free UDA without the need of source domain data, though still requiring a large amount of target domain data. Another setting is few-shot personalization, which utilizes a few labeled data samples to optimize the model performance for a specific person (Liu et al. 2019; Yu, Liu, and Odobez 2019; Chen and Shi 2020; Park et al. 2019). However, their experiments rely on gaze labels from off-the-shelf datasets, neglecting the difficulty to obtain such labels in practice. In this paper, we propose to address a more challenging problem, i.e., unsupervised personalization without source data.

2.2 Prompt Tuning for Computer Vision

Large foundation models have demonstrated exceptional effectiveness in natural language processing (NLP) and computer vision tasks. To fast adapt large models to downstream tasks, researchers have proposed continuous task- specific vectors that are updated via gradient, a method known as prompt tuning (Lester, Al-Rfou, and Constant 2021; Liu et al. 2023b, 2021a). In prompt tuning, the backbone parameters are fixed and only prompts are updated. As a parameter-efficient method, prompt tuning can achieve comparable results to full-parameter fine-tuning. In computer vision, prompting has initially been introduced to generalizing language models to address the few-shot and zero-shot classification problems (Radford et al. 2021). Despite its effectiveness, most works (Radford et al. 2021; Zhou et al. 2022; Ju et al. 2022; Yao et al. 2021) use both vision and language models and implement prompt by applying it only to the language model. Recently, (Jia et al. 2022) initially attempted to apply prompt tuning on pure vision models for few-shot learning. In (Jia et al. 2022), the authors propose two variants of prompt, i.e., additional inputs of vision transformer or tunable padding of convolutional layers. Following this work, (Kim, Kim, and Ro 2022, 2023) propose to use the prompting method in (Jia et al. 2022) for speaker-adaptive

speech recognition. In this paper, we follow (Jia et al. 2022) to apply tunable padding as prompt. In addition, we move one step forward to learn a meta-initialized prompt for better generalizing to a specific person's data.

2.3 Meta-Learning

Meta-learning, also known as learning-to-learn, has been widely applied in deep learning (Santoro et al. 2016; Snell, Swersky, and Zemel 2017; Finn, Abbeel, and Levine 2017; Zhong et al. 2022; Lian et al. 2019b). Meta-learning methods can be categorized as model-based (Santoro et al. 2016), metric-based (Snell, Swersky, and Zemel 2017) and optimization-based (Finn, Abbeel, and Levine 2017). An example is the model-agnostic meta-learning (MAML) (Finn, Abbeel, and Levine 2017) method. It aims to jointly learn a global optimal initialization of parameters and an update rule of task-specific parameters. As a result, the meta-learned model can generalize well to target tasks given a small number of samples with few gradient updates. Considering its ability to enable few-shot learning, (Park et al. 2019) propose to use meta-learning for few-shot gaze estimation. It proposes to use meta-learning to learn how to quickly adapt to unseen domains with few labeled data. Our use of meta-learning is different than that in (Park et al. 2019). Instead, our meta-learning approach is closer to a variant of MAML, i.e., meta-auxiliary learning (Liu, Davison, and Johns 2019; Chi et al. 2021; Liu et al. 2022, 2023a). It aims to address the primary task by solving an auxiliary task.

3 Problem Definition

In this paper, we focus on a rarely touched problem, i.e., test-time personalized gaze estimation. In contrast to unsupervised domain adaptation (UDA) for gaze estimation (Liu et al. 2021b; Wang et al. 2022; Bao et al. 2022), which emphasizes the average performance across all persons in the target domain, test-time personalization focuses on the performance with respect to a particular person in the target domain. To be specific, let $\mathcal{S} = \{(x_i^s, y_i^s)|x_i^s \in \mathcal{I}_S, y_i^s \in \mathcal{Y}_S\}_{i=1}^{N_S}$ denote the source dataset, where x_i^s and y_i^s are respectively the image and label from source image set $\mathcal{I}_S$ and source label set $\mathcal{Y}_S$. Similarly, let $\mathcal{A}_j = \{x_i^{a_j}|x_i^{a_j} \in \mathcal{I}_{T_j}\}_{i=1}^{N_{\mathcal{A}_j}}$ denote the personalization dataset of j-th person, where $x_i^{a_j}$ is the unlabeled image sampled from the target image set of j-th person $\mathcal{I}_{T_j}$. Our goal is to update the model f_θ learned on the source dataset S according to the personalization dataset $\mathcal{A}_j$, so that the resulting personalized model f_{θ_j} can perform better on the test data of the j-th person.

4 Preliminary of Source-Free UDA

We begin with the formulation of source-free unsupervised domain adaptation (UDA) before introducing our proposed approach. UDA methods (Bao et al. 2022; Liu et al. 2021b) consider a realistic problem that a model pre-trained on a dataset should be generalized to unseen data with distribution shift due to variations in subject appearance, lighting conditions, and image quality. We illustrate such variations in Figure 2 with examples from four datasets.

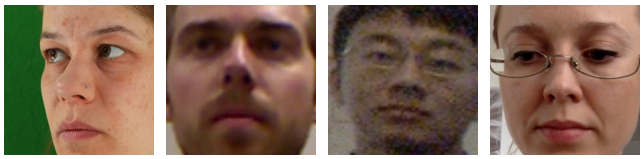

Figure 2: Illustration of the difference between four representative datasets. They are different from each other in subject appearance, image quality and lighting conditions.

UDA methods usually start from supervised pre-training on a source dataset $\mathcal{S}$. In the pre-training of gaze estimation on the source data, a typical practice (Bao et al. 2022; Liu et al. 2021b) is to learn a mapping f_θ via a combination of supervised and unsupervised losses:

$$\theta = \arg\min_\theta(\mathcal{L}_1(f_\theta(\mathcal{I}_S), \mathcal{Y}_S) + \mathcal{L}_{un}(f_\theta(\mathcal{I}_S), \mathcal{I}_S)), \quad (1)$$

where $\mathcal{L}_1$ is the supervised loss and $\mathcal{L}_{un}$ denotes the unsupervised loss. This pre-training phase aims to learn a model that can estimate gaze and accomplish unsupervised tasks, such as maintaining rotation consistency (Bao et al. 2022).

At the adaptation stage, only images $\mathcal{I}_T$ of the target domain are available for adapting the gaze estimation model. The domain adaptation of a typical UDA method, including (Liu et al. 2021b; Bao et al. 2022), can be summarized as:

$$\theta_j = \arg\min_\theta \mathcal{L}_{un}(f_\theta(\mathcal{I}_T), \mathcal{I}_T). \quad (2)$$

where the model parameters θ are fine-tuned by minimizing the unsupervised loss $\mathcal{L}_{un}$.

However, we argue that such a solution is not optimal in the context of personalizing a gaze estimation model where the target domain data from one person might be limited and similar among samples. First, optimizing a large number of parameters θ on such a small amount of data with similar patterns can easily lead to overfitting problem (Schneider and Vlachos 2021). Second, fine-tuning the entire network can be computationally costly, especially in scenarios with limited computing resources, such as adaptation on personal edge devices. In addition, although the unsupervised loss ($\mathcal{L}_{un}$) in domain adaptation has also been employed in the pre-training phase (see Equation 1), there is no explicit constraint ensuring that minimizing $\mathcal{L}_{un}$ is equivalent to minimizing $\mathcal{L}_1$. This can pose a risk of unsuccessful adaptation unless $\mathcal{L}_{un}$ is very carefully designed (Bao et al. 2022; Liu et al. 2021b).

5 Methodology

5.1 Prompting for Test-time Personalization

Unlike common machine learning approaches that learn a generic model for a broad user base, personalization acknowledges individual user characteristics, delivering a dedicated model for each. Test-time adaptation (Sun et al. 2020) is an potential way to enable personalization during testing with limited users' data. However, achieving test-time personalization without calibration or labels is still challenging.

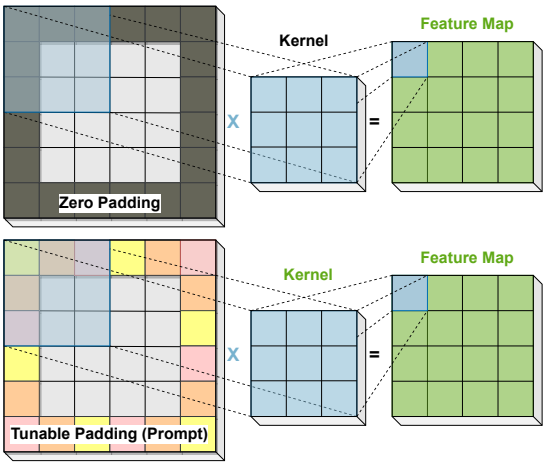

Figure 3: Illustration of replacing padding by prompt.

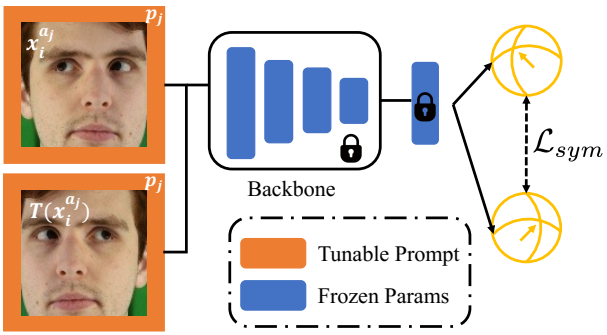

Figure 4: An overview of the proposed test-time personalization on j-th subject. In personalization, all the parameters are fixed except for the prompt.

Prompting for Personalized Gaze Estimation Inspired by the recent advances in NLP (Li and Liang 2021; Liu et al. 2023b; Lester, Al-Rfou, and Constant 2021), we incorporate the essence of the prompt tuning method into gaze estimation for efficient personalization. The key idea of prompting is to modify the inputs rather than network parameters (Li and Liang 2021). Following (Jia et al. 2022; Kim, Kim, and Ro 2022), we propose to transform the padding, an input to convolutional layers, to be our prompt. In convolutional layers, padding is usually employed to maintain or control the size of the output feature maps, such as zero padding and reflect padding. In these paddings, a predetermined number of zeros or reflections of the inputs are attached to the border of inputs before convolution operation. As shown in Figure 3, the padded region is then convolved with the kernel to produce the output feature map. In this way, a change in padding can impact the resulting feature embedding. Therefore, a natural idea is that we can modify the padding adaptively, guiding a network to produce desired feature embedding with respect to any specific person.

To instantiate this, we replace conventional padding with tunable parameters and update them during the personalization phase. By making the padding parameters trainable, we enable the network to learn the most appropriate padding strategy for a given subject. This, in essence, enables the network to be personalized.

Algorithm 1: Training of Meta Prompt

Require: Pre-trained network: f_θ
Require: Prompt: p
Require: learning rate λ_1 and λ_2
Output : meta-initialized prompt p

Initialize: Initialize the prompt p in f_θ by Gaussian noise and allow it to be trainable.
while *not converge* **do**
  Sample a mini-batch of training data in $\{\mathcal{I}_S, \mathcal{Y}_S\}$;
  for *each* x_i^s **do**
    Compute updated prompt $\hat{p}$:
    $\hat{p} = p - \lambda_1 \nabla_p \mathcal{L}_{per}(f_{\{\theta,p\}}(x_i^s), x_i^s)$
  Update:
  $p \leftarrow p - \lambda_2 \nabla_p \mathcal{L}_1(f_{\{\theta,\hat{p}\}}(x_i^s), y_i^s)$

It is worth mentioning that the number of parameters in prompt is far less than that in an entire network. For example, in ResNet-18, the prompt can contain less than 1% of the model parameters. The lightweight nature of the prompt makes it exceptionally suitable for test-time personalization. This can address the concerns raised in Section 4 about the overfitting problem and limited computational resources on edge devices.

Test-time Adaptive Gaze Estimation Given a model f_θ pre-trained on source data using Equation 1, we modify the model by incorporating a prompt (denoted as p), resulting in a new model $f_{\{\theta,p\}}$. At the test time, as discussed above, we update the prompt p instead of the entire set of model parameters practiced in existing methods (Liu et al. 2021b; Bao et al. 2022). We formalize the test-time training by rewriting the optimization problem in Equation 2 as follows:

$$p_j = \arg\min_p \mathcal{L}_{per}(f_{\{\theta,p\}}(\mathcal{A}_j), \mathcal{A}_j). \tag{3}$$

where p_j denotes the prompt corresponding to j-th person. $\mathcal{L}_{per}$ is an unsupervised loss used in both pre-training and personalization.

For the choice of $\mathcal{L}_{per}$, we want it to be gaze-relevant and computationally efficient. One option is the rotation consistency loss proposed in (Bao et al. 2022). Despite its effectiveness, the calculation of this loss demands multiple rotated versions of the original image as input to the network, resulting in great time cost in model adaptation. For ease of implementation and fast calculation, we adopt the left-right symmetry loss cited in (Kellnhofer et al. 2019). During the test time, it can be applied by the pre-trained model to regularize the gaze estimates for the specific individual being tracked. For a detailed description of the symmetry loss, please refer to the appendix. An overview of our test-time personalization is depicted in Figure 4.

5.2 Meta Prompt

As mentioned in Section 4, it is questionable that minimizing unsupervised losses $\mathcal{L}_{per}$ can naturally lead to a smaller gaze error, which is the ultimate goal. In this section, we present how meta-learning can align these two objectives. The recent method described in (Park et al. 2019) leverages meta-learning to enable rapid adaptation of gaze estimation model provided labels from the target domain. In this work, we explore the feasibility of constructing a "meta prompt" that can rapidly adapt to individual characteristics without the need for gaze labels. Inspired by the recent attempts in meta-learning (Chi et al. 2021; Liu, Davison, and Johns 2019; Liu et al. 2022) for image deblurring and classification, where the adaptation is performed via an auxiliary loss, we are further motivated to meta-initialize the prompt before tuning. The goal of the meta-training is to learn the prompt so that the gaze error is spontaneously minimized by optimizing the prompt based on the symmetry loss.

Given a pre-trained model f_θ, we randomly initialize the prompt p and obtain the model $f_{\{\theta,p\}}$ with the tunable prompt equipped. Note that the meta-training of prompt is conducted only on the source dataset $\mathcal{S}$. Next, we sample a mini-batch of N paired data $\{x_i^s, y_i^s\}_{i=1}^N$ and proceed to update the prompt for each sample using the personalization loss, as follows:

$$\hat{p} = p - \lambda_1 \nabla_p \mathcal{L}_{per}(f_{\{\theta,p\}}(x_i^s), x_i^s), \tag{4}$$

where λ_1 is the learning rate. Intuitively, this update allows the prompt to guide the network in generating features that reduce the left-right symmetry loss.

Recall that, our primary objective is to minimize the gaze error through updates to the prompt p. Accordingly, we aim to maximize the performance of the network by minimizing $\mathcal{L}_{per}$. We formally define our meta-objective as follows:

$$\arg\min_p \mathcal{L}_1(f_{\{\theta,\hat{p}\}}(x_i^s), y_i^s). \tag{5}$$

Note that the supervised loss is computed based on the network's result $f_{\{\theta,\hat{p}\}}(x_i^s)$, which is based on the updated prompt $\hat{p}$. However, the actual optimization is carried out on the prompt p. The meta-objective can be implemented by the following gradient descent:

$$p \leftarrow p - \lambda_2 \nabla_p \mathcal{L}_1(f_{\{\theta,\hat{p}\}}(x_i^s), y_i^s), \tag{6}$$

λ_2 being the learning rate.

As a summary, the above meta-training is designed to learn a general initialization of the prompt p that can be applied across individuals. Notably, since the subject being tested is unknown during the meta-learning phase, it is impossible to learn a personalized meta prompt p_j. The overall meta-learning procedure is summarized in Algorithm 1.

6 Experiments

6.1 Dataset

We employ four gaze estimation datasets as four different domains, namely ETH-XGaze ($\mathcal{D}_E$) (Zhang et al. 2020), Gaze360 ($\mathcal{D}_G$) (Kellnhofer et al. 2019), MPIIGaze ($\mathcal{D}_M$) (Zhang et al. 2017b), and EyeDiap ($\mathcal{D}_D$) (Funes Mora, Monay, and Odobez 2014). For preprocessing, we use the code provided in (Cheng et al. 2021). $\mathcal{D}_E$ and $\mathcal{D}_G$ are used as source domains whereas $\mathcal{D}_M$ and $\mathcal{D}_D$ are used as target domains, in alignment with RUDA (Bao et al. 2022) and PnP-GA (Liu et al. 2021b). Due to the page limit, please refer to supplementary material for the details of the datasets.

6.2 Implementation Details

Our method is implemented using PyTorch library (Paszke et al. 2019) and conducted on NVIDIA Tesla V100 GPUs. We use Adam (Kingma and Ba 2014) as our optimizer with $\beta = (0.5, 0.95)$. The training images are all cropped to a size of 224×224 without data augmentation.

Pre-training stage. During network pre-training, we use L1 loss and symmetry loss to train the network f_θ with a mini-batch size of 120. The initial learning rate is set to 10^{-4}. We train for 50 epochs with the learning rate multiplied by 0.1 at Epoch 25.

Meta-training stage. During the meta-training stage for prompt initialization, the network is initialized with weights obtained from the pre-training stage. Prompts are initialized randomly using a Gaussian distribution with mean 0 and variance 1. Note that if not explicitly specified, we replace the padding of the first nine convolutional layers in ResNet-18 (He et al. 2016). All other parameters are kept frozen. We use a mini-batch size of 20. The learning rates, i.e., λ_1 and λ_2 in Algorithm 1, are set to 10^{-4}. The meta-training process continues for 1000 iterations.

Personalization stage. To perform personalization for solving Equation 3, we use only 5 images per person. To ensure the reproducibility of results, we do not perform random sampling but use the first 5 images of each person. During personalization, only the prompt is optimized, and all other parameters are fixed. The learning rate is set to 0.01.

6.3 Comparison with the SOTA

To demonstrate the effectiveness and efficiency of our proposed method, we conduct a comparative study with several UDA methods for gaze estimation across four cross-domain tasks: $\mathcal{D}_E \rightarrow \mathcal{D}_M$, $\mathcal{D}_E \rightarrow \mathcal{D}_D$, $\mathcal{D}_G \rightarrow \mathcal{D}_M$, and $\mathcal{D}_G \rightarrow \mathcal{D}_D$. We select five representative methods, including source-available UDA (SA-UDA) and source-free UDA (SF-UDA) methods. Specifically, we compare our method with three SA-UDA methods, i.e., DAGEN (Guo et al. 2020), GazeAdv (Wang et al. 2019) and Gaze360 (Kellnhofer et al. 2019). For SF-UDA methods, we select PnP-GA (Liu et al. 2021b) and RUDA (Bao et al. 2022) for comparison. As SF-UDA methods can be easily adapted to our setting, we re-implemented these methods under our setting. For the re-implementations, we rely on their official code provided by the authors. For a fair comparison, the personalization of PnP-GA (Liu et al. 2021b) and RUDA (Bao et al. 2022) uses 10 images per person as the two methods are not specifically designed for personalization. ResNet-18 is used as the backbone for all methods.

Effectiveness of Our Method We present a comparison between the aforementioned methods in terms of gaze error in Table 1. We begin by comparing our method (TPGaze)

with a baseline based on supervised learning on the source dataset with L1 loss. Apparently, our method significantly outperforms the baseline in the cross-domain scenario.

Next, we compare our method with the SA-UDA approaches. All three SA-UDA approaches require labeled source data during the adaptation stage. Intuitively, these approaches can potentially achieve better performance due to the availability of source data. However, our method, which uses 5 personal images for adaptation only, significantly outperforms these SA-UDA methods by a large margin.

Finally, we compare with two SF-UDA methods. As previously noted, we re-implement these methods under our personalization setting. The last three rows of Table 1 clearly demonstrate that our method surpasses both PnP-GA and RUDA, even when those two methods are trained with 10 personal samples and employ unsupervised losses that are more complex than the symmetry loss we used. There can be two reasons: 1) Both PnP-GA and RUDA require all parameters to be tunable during the personalization stage. However, due to the limited amount of test domain data, this may potentially result in overfitting. In contrast, our method only allows a very small group of parameters to be tunable, mitigating the risk to overfit. 2) PnP-GA and RUDA assume that minimizing the outlier-guided loss or rotation consistency loss can be beneficial in reducing gaze error unconditionally. In comparison, our method proposes to use meta-learning to align the unsupervised loss with gaze error.

Efficiency of Our Method As stated in Section 5.1, efficient adaptation is a key advantage of our method. To prove this, we exhibit the number of tunable parameters for each method in the second column of Table 2. Our method requires only 0.125M parameters to be trainable during personalization, while RUDA requires at least 100 times more parameters to be updated. Additionally, special attention should be paid to PnP-GA, which relies on an ensemble of 10 networks and requires 116.9M parameters to be optimized for adaptation. This is 1000 times larger than our method. The excessive number of tunable parameters in the compared methods can potentially hinder their deployment on edge devices for online adaptation, while our method can be easily deployed on edge devices due to the fewer parameters requiring training.

Regarding the time cost of adaptation, although backpropagation is required throughout the entire model in our method, it is important to note that during the parameter update process, only 1% of a ResNet-18's parameters are updated. This selective parameter update can save time during adaptation. Besides, the symmetry loss used in our work can be fast calculated, unlike the one in RUDA that requires many times of image rotation. Here, we report the time cost of adaptation (duration/iteration) including loss calculation or not. By comparing the time solely for model updates (without loss calculation), we are 3 times faster than RUDA. And on the total time cost for adaptation (with loss calculation), we achieve a speed 10 times faster than RUDA, let alone the slower PnP-GA. The results further demonstrate the importance of prompt tuning and our loss selection.

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	8.02	7.30	7.79	8.19
DAGEN	7.53	8.46	9.31	12.05
GazeAdv	8.48	7.70	9.15	11.15
Gaze360	7.86	9.64	7.71	9.54
PnP-GA*	6.91	7.18	7.36	8.17
RUDA*	<u>6.86</u>	<u>6.84</u>	<u>6.96</u>	<u>5.32</u>
TPGaze	**6.30**	**5.89**	**6.62**	**5.04**

Table 1: Comparison with the state-of-the-art methods in terms of gaze error in degree, including source available UDA methods (middle rows) and our re-implementation (*) of source-free UDA methods under our personalization setting (bottom rows). The best results are in bold and the second best results are with underline.

Method	Tunable Params	Time w/ Loss	Time w/o Loss
PNP-GA	116.9M	0.390s	0.359s
RUDA	<u>12.20M</u>	<u>0.287s</u>	<u>0.089s</u>
TPGaze	**0.125M**	**0.029s**	**0.028s**

Table 2: Comparison in terms of adaptation efficiency (duration/iteration). Tunable parameters is the quantity of parameters that are updated during adaptation."Time w/ Loss" and "Time w/o Loss" respectively denote the time consumption in adaptation with or without loss calculation.

6.4 Ablation Study

We here conduct ablation studies to reveal the influence of the key components in our method. They are presented as follows: (a) **Baseline**: using ResNet-18 without personalization. (b) **Update All**: updating all parameters of the network (meta-learned) for personalization instead of prompt tuning. (c) **No Meta**: using $\mathcal{L}_{per}$ for personalization with randomly initialized prompts instead of meta prompts. (d) **TPGaze**: our proposed method.

The quantitative results of the ablation studies are shown in Table 3, demonstrating that our method with all components works the best. There are some other observations worth mentioning too. First, by comparing **TPGaze** with **Update All**, we find that prompt tuning is surprisingly better than tuning all parameters. This substantiates the effectiveness of introducing prompt for personalization. Second, by comparing **TPGaze** with **No Meta**, we validate the importance of using meta-learning for initializing prompts. Without meta-initialization, the performance of personalization drops. This highlights the significance of meta-learning in aligning the unsupervised loss with gaze error, leading to a more stable personalization process.

6.5 Additional Analysis

Influence of Prompt Size To study the influence of altering the number of layers with tunable prompt, we here conduct an additional analysis. The experimental results are shown in Table 4. We gradually increase the number from 0 (no prompt) to 17 (all convolutional layers). As the prompt

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	8.02	7.30	7.79	8.19
Update All	6.47	6.10	6.71	5.58
No Meta	6.59	6.18	6.68	5.76
TPGaze	**6.30**	**5.89**	**6.62**	**5.04**

Table 3: Ablation study of two components in our proposed method. Ours full solution (TPGaze) performs the best.

# Conv	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$	# Param
0	8.02	7.30	7.79	8.19	0
1	6.48	6.48	6.91	5.03	8.17K
5	6.38	6.29	6.67	**5.02**	66.5K
9	6.30	5.89	**6.62**	5.04	125.7K
13	**6.05**	**5.63**	6.96	5.45	186.8K
17	12.45	14.23	22.67	27.34	251.1K

Table 4: Influence of adding prompt to the different number of convolutional layers. "# Param" is the total number of parameters in the prompt.

# Samp	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$	Avg.
1	6.51	6.00	6.54	5.21	6.06
5	6.30	5.89	6.62	5.04	5.96
10	6.18	5.51	6.67	5.03	5.85
15	**6.14**	**5.26**	**6.54**	**5.02**	**5.74**

Table 5: The influence of using difference size of samples for personalization. We observe that in general more samples usually result in lower gaze error.

size increases, the average gaze error first decreases until the 9-th layer. Notably, the gaze errors when adding prompt to 9 layers and 13 layers are comparable, while the number of tunable parameters required for 13 layers is 48.9% higher than that of 9 layers. To maximize the benefits of accurate and efficient inference, we suggest incorporating prompts into 9 convolutional layers.

It is worth noting that when we add prompt to all the layers (17 convolutional layers), the performance decreases significantly. This could be attributed to the fact that the deepest prompt, located on the 17-th layer, can directly affect the output of the backbone. Since we do not update the final linear layer during personalization, the layer maintains the original mapping from the original backbone's outputs to the gaze direction. Thus, any major changes to the output features of the backbone may result in a defective mapping of the linear layer, leading to a catastrophic decrease in performance. This highlights the need for careful consideration when tuning prompts with respect to deeper layers.

Influence of Data Size As noted in Section 6.2, we only adopt 5 samples for the personalization. Here, we reveal the influence of using a different number of unlabeled samples for personalization. The results are presented in Table 5. Specifically, when we vary the number of samples from 1 to 15, we can observe that as the numberincreases, the gaze

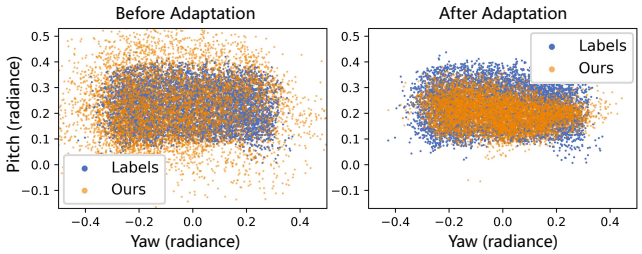

Figure 5: Distribution of gaze estimation results and ground-truth labels before and after personalization. Results are the personalization from $\mathcal{D}_G$ to $\mathcal{D}_D$.

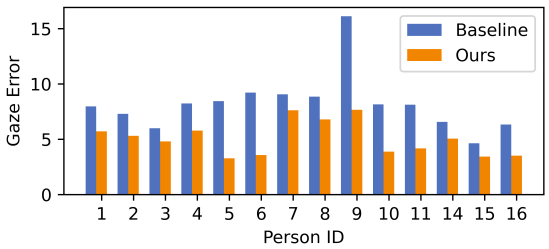

Figure 6: Breakdown of personalization results from $\mathcal{D}_G$ to $\mathcal{D}_D$.

errors tend to decrease, fitting our intuition that more unlabeled samples can potentially help improve the accuracy of the personalized gaze estimation model.

Visualization of Adaption Results To visualize the personalization results, we show the distribution of gaze estimation results before and after adaptation in Figure 5. It can be observed that the distribution of our approach (after adaptation) is considerably closer to the distribution of ground-truth labels, indicating a better adaptation performance.

Breakdown of the Personalization Results Due to the page limit, we primarily present the average gaze error across all individuals in a specific dataset. To provide further insight into our personalization results, we present a breakdown of the personalization results from $\mathcal{D}_G$ to $\mathcal{D}_D$ in Figure 6. It can be observed that our method can consistently outperform the baseline, as indicated by the lower gaze error.

7 Conclusion

In this paper, we present an efficient and accurate method to personalize gaze estimation at test time, without relying on labeled data. To achieve efficient personalization, we employ prompt tuning techniques. Moreover, we ensure that minimizing unsupervised loss aligns with minimizing gaze error through meta prompt. Our results show significant improvements against the state-of-the-art methods in terms of adaptation speed and accuracy. Besides, extensive additional analysis further demonstrates the strong performance and desirable properties of our proposed approach.

References

Admoni, H.; and Scassellati, B. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1): 25–63.

Bao, Y.; Liu, Y.; Wang, H.; and Lu, F. 2022. Generalizing gaze estimation with rotation consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4207–4216.

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

Burova, A.; Mäkelä, J.; Hakulinen, J.; Keskinen, T.; Heinonen, H.; Siltanen, S.; and Turunen, M. 2020. Utilizing VR and gaze tracking to develop AR solutions for industrial maintenance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–13.

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.

Castner, N.; Kuebler, T. C.; Scheiter, K.; Richter, J.; Eder, T.; Hüttig, F.; Keutel, C.; and Kasneci, E. 2020. Deep semantic gaze embedding and scanpath comparison for expertise classification during OPT viewing. In *ACM symposium on eye tracking research and applications*, 1–10.

Chen, Z.; and Shi, B. 2020. Offset calibration for appearance-based gaze estimation via gaze decomposition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 270–279.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 100–115.

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

Cheng, Y.; Zhang, X.; Lu, F.; and Sato, Y. 2020b. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29: 5259–5272.

Chi, Z.; Wang, Y.; Yu, Y.; and Tang, J. 2021. Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9137–9146.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135.

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

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

Ghosh, S.; Hayat, M.; Dhall, A.; and Knibbe, J. 2022. Mtgls: Multi-task gaze estimation with limited supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3223–3234.

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

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

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 709–727. Springer.

Ju, C.; Han, T.; Zheng, K.; Zhang, Y.; and Xie, W. 2022. Prompting visual-language models for efficient video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 105–124. Springer.

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, M.; Kim, H.; and Ro, Y. M. 2022. Speaker-adaptive Lip Reading with User-dependent Padding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 576–593. Springer.

Kim, M.; Kim, H.-I.; and Ro, Y. M. 2023. Prompt Tuning of Deep Neural Networks for Speaker-adaptive Visual Speech Recognition. *arXiv preprint arXiv:2302.08102*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; and Torralba, A. 2016. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2176–2184.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Lian, D.; Hu, L.; Luo, W.; Xu, Y.; Duan, L.; Yu, J.; and Gao, S. 2018. Multiview multitask gaze estimation with deep convolutional neural networks. *IEEE transactions on neural networks and learning systems*, 30(10): 3010–3023.

Lian, D.; Zhang, Z.; Luo, W.; Hu, L.; Wu, M.; Li, Z.; Yu, J.; and Gao, S. 2019a. RGBD based gaze estimation via multi-task CNN. In *Proceedings of the AAAI conference on artificial intelligence*.

Lian, D.; Zheng, Y.; Xu, Y.; Lu, Y.; Lin, L.; Zhao, P.; Huang, J.; and Gao, S. 2019b. Towards fast adaptation of neural architectures with meta learning. In *International Conference on Learning Representations*.

Liu, G.; Yu, Y.; Mora, K. A. F.; and Odobez, J.-M. 2019. A differential approach for gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 43(3).

Liu, H.; Chi, Z.; Yu, Y.; Wang, Y.; Chen, J.; and Tang, J. 2023a. Meta-Auxiliary Learning for Future Depth Prediction in Videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5756–5765.

Liu, H.; Wu, Z.; Li, L.; Salehkalaibar, S.; Chen, J.; and Wang, K. 2022. Towards Multi-Domain Single Image Dehazing via Test-Time Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5831–5840.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Liu, S.; Davison, A.; and Johns, E. 2019. Self-supervised generalisation with meta auxiliary learning. *Advances in Neural Information Processing Systems*, 32.

Liu, X.; Ji, K.; Fu, Y.; Du, Z.; Yang, Z.; and Tang, J. 2021a. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Liu, Y.; Liu, R.; Wang, H.; and Lu, F. 2021b. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3835–3844.

Park, S.; Mello, S. D.; Molchanov, P.; Iqbal, U.; Hilliges, O.; and Kautz, J. 2019. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9368–9377.

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

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Reale, M.; Hung, T.; and Yin, L. 2010. Viewing direction estimation based on 3D eyeball construction for HRI. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, 24–31. IEEE.

Ruzzi, A.; Shi, X.; Wang, X.; Li, G.; De Mello, S.; Chang, H. J.; Zhang, X.; and Hilliges, O. 2023. GazeNeRF: 3D-Aware Gaze Redirection with Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9676–9685.

Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; and Lillicrap, T. 2016. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, 1842–1850.

Schneider, J.; and Vlachos, M. 2021. Personalization of deep learning. In *Proceedings of the 3rd International Data Science Conference–iDSC2020*, 89–96. Springer.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A.; and Hardt, M. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, 9229–9248.

Wang, K.; Zhao, R.; Su, H.; and Ji, Q. 2019. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11907–11916.

Wang, Y.; Jiang, Y.; Li, J.; Ni, B.; Dai, W.; Li, C.; Xiong, H.; and Li, T. 2022. Contrastive regression for domain adaptation on gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19376–19385.

Wollaston, W. H. 1824. Xiii. on the apparent direction of eyes in a portrait. *Philosophical Transactions of the Royal Society of London*.

Wood, E.; and Bulling, A. 2014. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the symposium on eye tracking research and applications*.

Yao, Y.; Zhang, A.; Zhang, Z.; Liu, Z.; Chua, T.-S.; and Sun, M. 2021. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*.

Yu, Y.; Liu, G.; and Odobez, J.-M. 2019. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11937–11946.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 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.

Zhong, T.; Chi, Z.; Gu, L.; Wang, Y.; Yu, Y.; and Tang, J. 2022. Meta-DMoE: Adapting to Domain Shift by Meta-Distillation from Mixture-of-Experts. In *Advances in Neural Information Processing Systems*, volume 35, 22243–22257.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.