ELF-UA: Efficient Label-Free User Adaptation in Gaze Estimation

Yong Wu[1] , **Yang Wang**[2] , **Sanqing Qu**[1] , **Zhijun Li**[1] and **Guang Chen**[1*]
[1]Tongji University
[2]Concordia University
{yongwu, 2011444, guangchen}@tongji.edu.cn, yang.wang@concordia.ca, zjli@ieee.org

Abstract

We consider the problem of user-adaptive 3D gaze estimation. The performance of person-independent gaze estimation is limited due to interpersonal anatomical differences. Our goal is to provide a personalized gaze estimation model specifically adapted to a target user. Previous work on user-adaptive gaze estimation requires some labeled images of the target person data to fine-tune the model at test time. However, this can be unrealistic in real-world applications, since it is cumbersome for an end-user to provide labeled images. In addition, previous work requires the training data to have both gaze labels and person IDs. This data requirement makes it infeasible to use some of the available data. To tackle these challenges, this paper proposes a new problem called efficient label-free user adaptation in gaze estimation. Our model only needs a few unlabeled images of a target user for the model adaptation. During offline training, we have some labeled source data without person IDs and some unlabeled person-specific data. Our proposed method uses a meta-learning approach to learn how to adapt to a new user with only a few unlabeled images. Our key technical innovation is to use a generalization bound from domain adaptation to define the loss function in meta-learning, so that our method can effectively make use of both the labeled source data and the unlabeled person-specific data during training. Extensive experiments validate the effectiveness of our method on several challenging benchmarks.

1 Introduction

Gaze is a crucial non-verbal cue for understanding the internal states of humans. Gaze estimation from images has been widely used in various applications, such as saliency detection [Wang *et al.*, 2020], human-computer interaction [Zhang *et al.*, 2017a], and virtual reality industry [Prange *et al.*, 2017; Chen *et al.*, 2020; Konrad *et al.*, 2020]. In this paper, we focus on predicting 3D gazes since it has more applications than 2D gazes. Most existing gaze estimation approaches [Zhang *et al.*, 2017a; Zhang *et al.*, 2017b; Krafka *et al.*, 2016] use standard supervised learning to learn a person-independent model. As observed in [Park *et al.*, 2019], the performance of person-independent gaze estimation is limited due to interpersonal anatomical differences. In order to achieve the high quality required by many applications, it is much more desirable to personalize the model to each user. Unfortunately, it is unrealistic to train a personalized gaze estimation for each user using standard supervised learning since it is impossible to collect enough training data for each user. Some recent work [Park *et al.*, 2019] proposes a setting called few-shot adaptive gaze estimation. It uses a meta-learning approach to learn a model during offline training (called "meta-training"

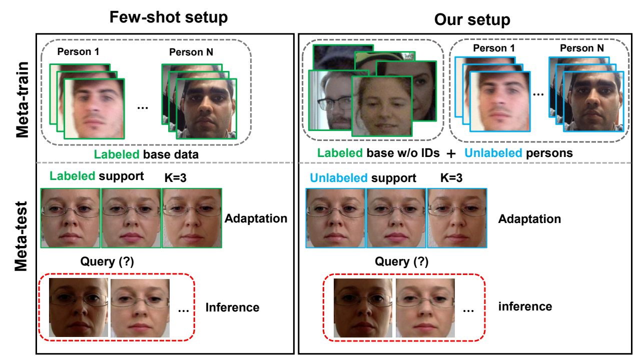

Figure 1: Illustration of our problem setup. (**Left**) In previous work on few-shot user adaptive gaze estimation [Park *et al.*, 2019], the model is trained on the labeled base dataset with many persons (also called *tasks*) during meta-training. During meta-testing, given a few labeled samples from a new person (known as *support set*), the model is adapted to the new person. Then the model predicts other images (called *query set*) from this person. (**Right**) In this work, we propose a new problem setting called label-free user adaptation in gaze estimation. During meta-training, we have a labeled source dataset without person IDs. We also have some unlabeled data with person IDs. During met-testing, we are provided very few *unlabeled* images ($\leq$5) from a new target person. Our goal is to get a gaze model specifically adapted to this target person. Note that our problem setup does not require any labeled images from the target person for adaptation. Our problem setting is closer to real-world scenarios.

[*]Corresponding author

in meta-learning). During testing (called "meta-testing" in meta-learning), the model is adapted to a new user with a few labeled images of this user.

Compared with standard supervised learning, the few-shot adaptive gaze estimation setting in [Park *et al.*, 2019] represents a significant advancement toward real-world use. However, this setting still has limitations that prevent practical deployment. First, during meta-testing, it requires a few *labeled* images of the target user in order to adapt the model to a new user. In practice, it is unrealistic to ask users to provide annotations in real-world applications. This is especially for 3D gaze, since the 3D gaze annotation may require specialized hardware (e.g. Tobii eye tracker) inaccessible to regular users. In contrast, it is much easier for the end user to simply capture a small number (e.g. 1 to 5) of unlabeled images. It is much more desirable if we could adapt the model using only a few unlabeled images of the target user. Second, during meta-training, the work in [Park *et al.*, 2019] requires the training data to be collected from different persons, where each person has multiple images. Each training image has to be fully annotated with both the gaze label and the person ID. The person ID is needed to construct the tasks in meta-learning, where each task corresponds to a user. This puts too much constraint on the training data. On one hand, some existing gaze estimation datasets [Kellnhofer *et al.*, 2019] do not come with person IDs. The method in [Park *et al.*, 2019] cannot make use of these datasets. For some large companies (e.g. Google, Amazon and Baidu), they are easy to collect many unlabeled face images without person IDs. However, the annotations are expensive. On the other hand, many 3D gaze datasets are collected in controlled environments [Zhang *et al.*, 2020; Funes Mora *et al.*, 2014] due to the difficulty of obtaining ground-truth 3D gaze labels. For training 3D gaze estimation models for real-world applications, we eventually have to collect new data to cover diverse real-world scenarios (e.g. users interacting with cell phones, tablets, smart TVs, etc). It can be difficult to collect labeled images since these images might be crowd-sourced from end users. Unlike other computer vision tasks (e.g. image classification) where ground-truth labels can be annotated by humans in a post hoc way, the 3D gaze label has to be obtained from specialized hardware (e.g. Tobii eye tracker) at the time when the image itself is taken. Due to the domain shift between existing gaze datasets and different users in real-world scenarios, it is not optimal to directly deploy a gaze model learned from existing labeled gaze datasets to a new user. In order to enable new applications of gaze estimation in the future, we need to come up with a way of making use of the diverse data even if they are unlabeled.

To address the aforementioned issues, we propose a new problem setting called efficient label-free user adaptation in gaze estimation. As shown in Fig. 1, there are two key differences between our setting and the few-shot setting in [Park *et al.*, 2019]. First of all, the work in [Park *et al.*, 2019] requires a small number (e.g. 1 to 5) of *labeled* images to adapt to a new user during meta-testing. In contrast, we only need a small number (e.g. 1 to 5) of *unlabeled* images for the model adaptation. Second, all training data in [Park *et al.*, 2019] have to be annotated with both gaze labels and person IDs. In contrast, our method only requires a source dataset annotated with gaze labels (without person ID) and some unlabeled person-specific dataset with person ID (but without gaze labels). In practice, the source data can be any existing gaze dataset. The unlabeled person-specific data can be easily obtained by crowd-sourcing from end users without requiring specialized hardware or manual annotation. Table 1 highlights the differences between our problem setting and other related problem settings, including unsupervised domain adaptation (UDA) [Cai *et al.*, 2020; Liu *et al.*, 2021], source-free domain adaptation (source-free DA) [Liang *et al.*, 2020] and few-shot learning (FSL) [Park *et al.*, 2019].

Our proposed method is based on meta-learning, in particular model-agnostic meta-learning (MAML) [Finn *et al.*, 2017]. Our network architecture consists of a main branch for the gaze prediction and an auxiliary branch for a self-supervised task with a shared backbone. Each user is considered a task in MAML. During meta-testing, we adapt the model to a target user using only a few unlabeled images by optimizing the loss function of the self-supervised auxiliary task. During meta-training, the model is trained to facilitate effective fast adaptation to each user using a bi-level optimization. The main challenge is that since the person-specific images during meta-training are unlabeled, we cannot easily define a loss for the outer loop of MAML as in [Park *et al.*, 2019; Finn *et al.*, 2017]. The key innovation of our work is to introduce a surrogate loss based on a theoretical result from domain adaptation for the outer loop in MAML.

Summary of Contributions: The contributions of this work are manifold. First, we present a new problem called efficient label-free user adaptation in gaze estimation. Different from the standard supervised learning [Zhang *et al.*, 2017a] or adaptation adaptation setup [Park *et al.*, 2019; Yu *et al.*, 2019], our problem formulation for training uses some labeled source data without person IDs and some unlabeled person-specific data, which learns a better person-specific gaze model. For an unseen person, the trained model only uses unlabeled samples to adapt the gaze model for this person. Second, we propose a novel approach to address this problem, which uses generalization bound from domain adaptation to define the loss function in meta-learning. Our model can effectively adapt to a new person given the a very few unlabeled images from that person at test time. Finally, we conduct an extensive evaluation of the proposed approach on several challenging benchmark datasets. Our approach significantly outperforms other alternatives and achieves competitive performance compared with other state-of-the-art methods.

2 Related Work

In this section, we review prior work in four main lines of research most relevant to our work, namely gaze estimation, domain adaptation, meta-learning and self-supervised learning.

Gaze Estimation. Most previous work in gaze estimation users supervised learning. These methods require large-scale labeled data. However, it is hard to obtain a large amount of

Method	Training		Adaptation				Inference
	Source data	Source labels	Source data	Target data	Target labels	Efficient	Model
UDA [Liu *et al.*, 2021]	■■■■	■■■■	✓	■■■■	✗	No	Dataset-specfic
Source-free DA [Qu *et al.*, 2022]	■■■■	■■■■	✗	■■■■	✗	No	Dataset-specific
FSL [Park *et al.*, 2019]	■■■■	■■■■	✗	■	✓	Yes	Person-specific
Ours	■■■■	■■	✗	■	✗	Yes	Person-specific

Table 1: Comparison of different problem settings in gaze estimation. If a method can achieve its objective with only a few gradient updates using a small amount of data, we consider it to be efficient. We also distinguish whether the adapted model is designed for a specific dataset or for a specific individual. Our proposed approach only requires a few unlabeled images of a user for adaptation. Our approach is efficient since it only needs a few gradient updates for adaptation. Furthermore, our approach can effectively utilize both labeled and unlabeled data during training, as long as the unlabeled data includes person IDs.

labeled data, especially for 3D gaze which requires specified hardware (e.g. Tobii eye tracker) to obtain the ground-truth label. In addition , due to the different anatomical structures (e.g. eye shape, visual axis) between different persons, the performance of supervised gaze estimation models is limited when testing on unseen persons. Some recent work [Lu *et al.*, 2014; Park *et al.*, 2019; Lindén *et al.*, 2019; Yu *et al.*, 2019; Liu *et al.*, 2024] propose to build personalized gaze estimation specifically for each target user. For example, finetuning a pretrained model [He *et al.*, 2019; Yu *et al.*, 2019; Park *et al.*, 2019] can achieve very high accuracy for a specific person, but it usually requires annotated data from this person during testing. This requirement leads to difficulties in real-world applications. In our work, we aim to address this issue with label-free user adaptation where the adaptation only requires a few unlabeled images of the target user.

Domain Adaptation. Domain adaptation (DA) aims to alleviate the distribution differences between the two domains by transferring the knowledge from the source domain to the target domain [Liu *et al.*, 2021; Cai *et al.*, 2020; Liang *et al.*, 2020; Qu *et al.*, 2022; Cai *et al.*, 2019; Xu *et al.*, 2020; Chi *et al.*, 2024; Wu *et al.*, 2024; Peng *et al.*, 2024]. It can minimize statistical discrepancy across domains. For gaze estimation, there is often a domain shift between datasets due to differences in terms of image capture devices and environments. There is also domain shift between different users due to inter-personal anatomical difference. Adversarial [Cui *et al.*, 2020; Long *et al.*, 2018] learning is applied to explore unaligned feature space between source dataset and target dataset. However it has access to target samples in advance of inference. Unsupervised domain adaptation (UDA) [Liu *et al.*, 2021] is another direction of research to alleviate domain shift, which has no access to target labels during training. It also needs target samples. Few-shot learning (FSL) [Park *et al.*, 2019; Yu *et al.*, 2019] is also adopted to improve performance when there is indistinguishable between different domains. Nevertheless, FSL needs a large number of labeled person-specific data to train the model during meta-training, then update the model with few labeled target samples during meta-testing. It is unrealistic in a real-world, since the acquisition of labeled data during testing is very difficult. In contrast, our method only requires a labeled source dataset without person ID and unlabeled person-specific dataset to train a personalized gaze model. The source dataset can be any existing gaze dataset or collected from web. Further, our method uses a few unlabeled target samples to update the trained model for each unseen person at test time.

Meta-Learning. Recently, meta-learning [Shen *et al.*, 2021; Finn *et al.*, 2017; Sung *et al.*, 2018; Lee *et al.*, 2019; Vinyals *et al.*, 2016; Snell *et al.*, 2017; Wu *et al.*, 2023; Zhong *et al.*, 2022; Chi *et al.*, 2022] has been explored for quick adaptation to new categories. Meta-learning can learn transferable information for various tasks, which is critical in few-shot learning. More specifically, the existing meta-learning methods can be categorized as the metric-based [Snell *et al.*, 2017; Sung *et al.*, 2018], model-based [Munkhdalai and Yu, 2017] and optimization-based [Finn *et al.*, 2017; Nichol *et al.*, 2018]. Our proposed approach builds upon one of the most popular meta-learning algorithms, namely model agnostic meta-learning (MAML) [Finn *et al.*, 2017]. MAML involves a bi-level optimization, including inner loop and outer loop. The goal of MAML is to learn a model initialization such that it can be quickly adapted to any new task with a few labeled data. Different from standard MAML, our proposed approach does not require labels of tasks during meta-training, and does not require any labeled data for adaptation.

Self-Supervised Learning. Self-supervised learning [Chi *et al.*, 2021; Carlucci *et al.*, 2019; Gidaris *et al.*, 2018; Tang *et al.*, 2021; Liang *et al.*, 2022; An *et al.*, 2021] is often used to learn feature representations by solving a pretext task, where the label of the pretext task can be obtained from an image itself. In computer vision, many self-supervised pretext tasks have been proposed, such as rotation degrees [Gidaris *et al.*, 2018] or shuffling order of patch-shuffled images [Carlucci *et al.*, 2019]. Recently, there have some advanced self-supervised learning approaches using contrastive learning. Typically, self-supervised approaches learn generic feature representations which mitigate over-fitting to domain-specific biases. Carlucci *et al.* [Carlucci *et al.*, 2019] propose to solve a jigsaw puzzle as the auxiliary task, where the image is cropped the shuffled image patches and the model is trained to learn the right orders. This helps the model to learn the concepts of spatial correlation while acting as a regularizer for the classification task. In this work, we use the jigsaw puzzle as a self-supervised task for model adaptation in gaze estimation. Since the whole face images provide rich gaze information [Cheng *et al.*, 2020; Wu *et al.*, 2022] than eyes. Reorder the face image at the training allows the model to learn internal gaze information.

3 Preliminaries

In this section, we first describe our problem setting, then we introduce a generalization bound that is the foundation of many domain adaptation approaches. Our proposed method builds upon this generalization bound.

3.1 Problem Setting

During offline training (called "meta-training" using meta-learning terminology), we are given a labeled source dataset $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N_s}$, where x_i is a training image and y_i is the corresponding gaze label. The instances of the source data are *not* person specific. In other words, we do not know the person ID of each instance (e.g. dataset Gaze360 [Kellnhofer *et al.*, 2019]). It is possible for each instance in $\mathcal{S}$ to be of a unique person. The standard supervised gaze estimation usually trains a model $f_\theta(\cdot)$ with parameter θ directly using $\mathcal{S}$. This model is then used during testing for any user.

In addition to the source dataset $\mathcal{S}$, we also have access to an *unlabeled person-specific* dataset $\mathcal{T}$. This person-specific dataset contains unlabeled images of K persons, where each person has multiple images, i.e. $\mathcal{T} = \{\mathcal{D}_1, \mathcal{D}_2, ...\mathcal{D}_K\}$ where $\mathcal{D}_i = \{x_j\}_{j=1}^{N_i}$ is the subset of N_i images corresponding to the i-th person. We assume N_i is reasonably large In practice, this person-specific dataset can be fairly easily captured using smartphones or web cameras. We assume that there is a domain shift between the source dataset and each user, i.e. $P_\mathcal{S} \neq P_{\mathcal{D}_i}$. There is also a domain shift between two different users, i.e. $P_{\mathcal{D}_i} \neq P_{\mathcal{D}_j}$ for $i \neq j$. Due to this domain shift, a model learned from $\mathcal{S}$ will not work well for a particular person $\mathcal{D}_i$. Ideally, we would like to have a personalized model for each user.

During testing (called "meta-testing" using meta-learning terminology), we have a new user. For this user, we only have *few-shot unlabeled* images $\mathcal{D} = \{x_j\}_{j=1}^{d}$ where d is small (e.g. 1 to 5). Our goal is to adapt the model θ using $\mathcal{D}$ and obtain an adapted model θ'. Then the adapted model θ' will be used for prediction for new images of this user.

Remark. In the real-world scenarios, many systems tuned to specific environment or person need to be calibrated. we can easily collect some images via web cameras (collecting many enough images are still challenging), but acquisition of labels is very difficult. This leads to inferior performance even if the system is calibrated by unlabeled images. Motivated by this, we propose this new problem setting, which pans out for this problem. We can update the system tailored to a new environment or person with a very few samples. Existing methods in the literature cannot address this new problem setting. Standard few-shot setting in [Park *et al.*, 2019; Yu *et al.*, 2019] needs labeled images of target persons for adaptation, while unsupervised domain adaptation or source-free domain methods [Cai *et al.*, 2020; Cui *et al.*, 2020] require target person images to align the model during training.

3.2 Generalization Bound in Domain Adaptation

In this section, we briefly introduce a mathematical tool in analyzing generalization bound for domain adaptation [Mansour *et al.*, 2009; Ben-David *et al.*, 2010]. Our proposed method builds upon this generalization bound.

Let $\mathcal{S}$ be a source domain and $\mathcal{T}$ be a target domain. There is a domain shift between these two domains. Given a machine learning model f_θ that maps an input x to an output y, $L(f_\theta(x), y)$ be a loss function that measures the difference between the prediction $f_\theta(x)$ and the ground-truth y. Then the following bound holds:

Theorem 1. $$\mathbb{E}_\mathcal{T}[L(f_\theta(x), y)] \leq \mathbb{E}_\mathcal{S}[L(f_\theta(x), y)] + d(\mathcal{S}, \mathcal{T}), \tag{1}$$

where $\mathbb{E}_\mathcal{S}[\cdot]$ (or $\mathbb{E}_\mathcal{T}[\cdot]$) denotes the expectation with respect to the distribution of the domain $\mathcal{S}$ (or $\mathcal{T}$). Here $d(\mathcal{S}, \mathcal{T})$ is a measure that quantifies the difference between $\mathcal{S}$ and $\mathcal{T}$. In practice, $d(\cdot)$ is defined such that it does not require labels in the target domain.

Theorem 1 provides the theoretical foundation for many domain adaptation methods. In domain adaptation, we would ideally like to minimize the loss $\mathbb{E}_\mathcal{T}[L(f_\theta(x), y)]$ on the target domain. However, this loss is not well defined on the target domain since we only have unlabeled data in the target domain. Instead, domain adaptation methods usually optimize the upper bound on the right hand side of Theorem 1. Maximum Mean Discrepancy (MMD) [Long *et al.*, 2017] is a popular method to measure the distance between different domains, and we can also use it to solve $d(\mathcal{S}, \mathcal{T})$ in Theorem 1. Note that in order to use MMD for domain adaption, we need to know the target domain in advance, and the unlabeled data in the target domain should be large enough.

4 Our Approach

We propose a novel meta-learning framework for learning self-supervised user-adaptive gaze estimation. Our proposed framework is based on the model-agnostic meta learning (MAML) [Finn *et al.*, 2017]. The architecture overview is illustrated in Fig. 2. During meta-training, we learn to transfer useful knowledge from the labeled source dataset $\mathcal{S}$ and the unlabeled person-specific dataset $\mathcal{T}$ to achieve generalization on unseen persons. We consider each person in $\mathcal{T}$ as a task in meta-learning. For each task, we construct a task with a support set D^{tr} and query set D^{val}. The support set D^{tr} is used to update the model parameter ψ to obtain a user-adapted model ψ'. The query set D^{val} is used to measure the performance of the adapted model ψ'. The meta-training procedure involves bi-level optimization as shown in Fig. 2, where the inner loop performs task-level optimization, while the outer loop performs a global model update via meta-objective.

Self-Supervised Inner Update. In the inner loop of meta-training, we are given a sampled task (i.e. person) i. We use D_i^{tr} to denote the support set of this task. Different from [Park *et al.*, 2019], our support set D_i^{tr} only contains a few unlabeled images of this person. Our goal is to update the model ψ using D_i^{tr} to obtain a user-adapted model ψ_i'.

Inspired by [Sun *et al.*, 2020], we add an additional head to our network for a self-supervision auxiliary task. Fig. 2 shows the illustration of the proposed model. The face images are decomposed to 16 patches with a 4×4 grid, which are randomly shuffled and then recomposed the face images. We define a set of M patch permutations and assign an index to each of them. The original ordered and the shuffled images are passed to the network. The auxiliary task is to predict

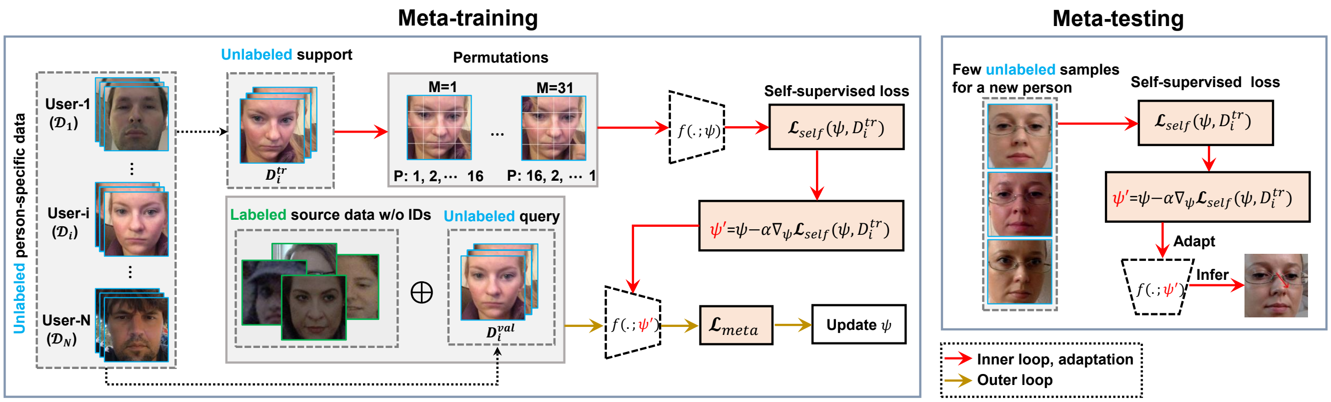

Figure 2: Overview of our approach. Our training data consist of a source dataset $\mathcal{S}$ annotated with gaze labels but without person IDs. We also have an unlabeled subject-specific dataset, where images are annotated with person IDs but without gaze labels. Each subject corresponds to a "task" in meta-learning. For each task i, we construct its support set D_i^{tr} and its query set D_i^{val}, both of which consist of unlabeled images. In the inner loop of meta-training, we use a self-supervised auxiliary task (permutation prediction) defined on the support set D_i^{tr} to update the model ψ to a user-adapted model ψ'. Since the query set D_i^{val} is unlabeled, we cannot directly compute a supervised loss of ψ' on D_i^{val}. Instead, we use a domain adaptation loss defined on $\mathcal{S}$ and D_i^{val} as an upper bound approximation to the supervised loss on D_i^{val}. This domain adaptation loss is used for the outer loop of meta-training. After meta-training, the model has learned to effectively adapt to a new subject using only a few unlabeled images. During meta-testing, we adapt the model to a new subject and use the adapted model for inference.

the permutation from the shuffled image. The face images enable to provide more gaze cues [Cheng *et al.*, 2020; Wu *et al.*, 2022]. During the inner loop, the overall self-supervised loss on the support set is defined as:

$$\mathcal{L}_{self}(\psi; D_i^{tr}) = \sum_{x\in D_i^{tr}} \ell_M(f(z_k|x), M_k). \quad (2)$$

Note that z_k is the index for the input shuffled image, and M_k is the index for the original order. ℓ_M is the standard cross-entropy loss for predicting the right order. During the inner loop, we do not need any labels for training the model, which is able to exploit the internal gaze information by solving the auxiliary self-supervised task.

Our goal is to update the model parameters ψ to obtain a person-adaptive model ψ_i' specifically tuned to the task $\mathcal{D}_i$. We obtain ψ_i' by taking a small number of gradient updates using Eq. 2 as:

$$\psi_i' \leftarrow \psi - \alpha\nabla_\psi\mathcal{L}_{self}(\psi; D_i^{tr}). \quad (3)$$

Domain Adaptation Based Outer Update. The adapted model ψ_i' is specifically tuned to the i-th person. Intuitively, we would like ψ_i' to perform well on other images of the same person. In previous applications of MAML [Finn *et al.*, 2017; Park *et al.*, 2019], one can uses the loss defined on the query set D_i^{val} as this performance measure. However, since D_i^{val} only contains unlabeled images in our setting, it is not trivial how to directly define a loss using D_i^{val}.

Our key insight is that we can approximate the supervised loss defined on D_i^{val} by its upper bound provided by RHS of Theorem 1. We can measure the performance of ψ_i' on D_i^{val} of $\mathcal{D}$ as:

$$\mathcal{L}_{meta}(\psi_i'; D_i^{val}) = \mathcal{L}_{gaze}(\psi_i'; \mathcal{S}) + \gamma\cdot d(\mathcal{S}, D_i^{val}; \psi_i'), \quad (4)$$

where $\mathcal{L}_{gaze}(\psi_i'; \mathcal{S})$ is the standard supervised loss for gaze prediction defined on the labeled source data $\mathcal{S}$. Here we use the angular gaze estimation error in terms of L1 loss for $\mathcal{L}_{gaze}$. Here $d(\mathcal{S}, D_i^{val}; \psi_i')$ is a measure of the difference between two domains $\mathcal{S}$ and $\mathcal{D}_i^{val}$. Inspired by [Long *et al.*, 2017], we implement $d(\mathcal{S}, D_i^{val}; \psi_i')$ using joint maximum mean discrepancy (joint MMD) as:

$$d(S, D_i^{val}, \psi_i') = ||\mathcal{C}_{Z^{1:n}}(S) - \mathcal{C}_{Z^{1:n}}(D_i^{val})||^2, \quad (5)$$

here n is meta-batch size. The $\mathcal{C}_{Z^{1:n}}$ is empirical joint embedding. For unlabeled person-specific data D_i^{val}, it can be estimated as $\mathcal{C}_{Z^{1:n}}(D_i^{val}) = \frac{1}{t}\sum_{j=1}^{t}\otimes_{l=1}^{n}\phi^l(\mu_j^l)$, where t is number of samples in D_i^{val}, and μ_j is a sample drawn from D_i^{val}. It is noted joint MMD does not require labels, but it implicitly depends on ψ_i' since it is defined on the features extracted by the backbone network.

The goal of meta-learning is to learn a good initial model ψ, so that after model update using the support set (Eq. 3) of a particular task, the task-adapted

model ψ_i' will minimize the loss defined in Eq. 4 on the corresponding query set across K tasks (persons). The meta-objective is defined as:

$$\min_\psi \sum_{i=1}^{K} \mathcal{L}_{meta}(\psi_i'; D_i^{val}). \quad (6)$$

The initial model ψ is learned by optimizing the meta-objective (Eq. 6) using stochastic gradient descent as:

$$\psi \leftarrow \psi - \beta\nabla_\psi\sum_{i=1}^{N}\mathcal{L}_{meta}(\psi_i'; D_i^{val}), \quad (7)$$

where β represents the meta-learning rate. The meta-objective in Eq. 6 involves summing over all meta-training

Algorithm 1 Training for ELF-UA

Require: Distribution $p(\mathcal{T})$ over tasks $\mathcal{D}$; labeled source dataset P_S;
Require: Learning rates α and β; meta-batch size n
Initialize network θ
while not done **do**
  Sample K unlabeled tasks (i.e. persons) $\mathcal{D}_i \sim p(\mathcal{T})$
  for all $\mathcal{D}_i$ **do**
    Construct $\mathcal{D}_i = \{D_i^{tr}, D_i^{val}\}$
    Construct self-supervised loss $\mathcal{L}_{self}(\theta; D_i^{tr})$
    Obtain the adapted model ψ_i' via Eq. 3 on D_i^{tr}
    Construct upper bound of outer loop loss $\mathcal{L}_{meta}(\psi_i'; D_i^{val})$
    Evaluate meta-objective in Eq. 4 on D_i^{val}
  end for
  Update model parameters ψ via Eq. 7
end while

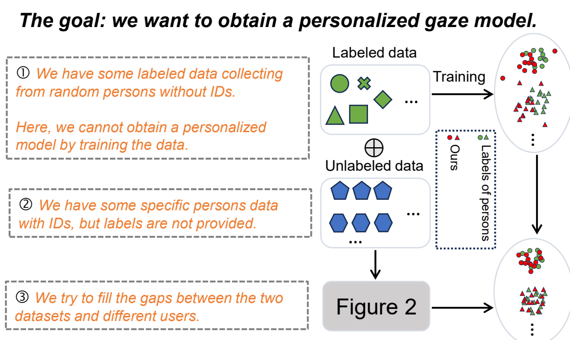

Figure 3: The flow of the proposed method.

tasks. In practice, we sample a mini-batch of tasks (persons) in each iteration during meta-training and sum over these sampled tasks. The entire training procedure is elaborated in Algorithm 1 and Fig. 2. We also give the flow of the proposed method in Fig. 3.

Meta-Testing (Adaptation + Inference). After meta-training, the initial model parameters ψ have been learned to effectively adapt to a person using only a small number of unlabeled images. During testing (also called meta-testing in our work which consists of adaptation and inference stages), given a new person $\mathcal{D} = (D^{tr}, D^{val})$ with the support set D^{tr} and the query set D^{val}. We simply use Eq. 3 to obtain the person-specific parameters ψ' via the self-supervised loss. The ψ' is adopted for gaze estimation in other face images from this person. Note that we do not need any labels for the adaptation during meta-testing.

5 Experiments

In this section, we first introduce the datasets and setup. We then describe our implementation details. We present our experimental results and comparisons.

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_{GC}$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_{GC}$
Baseline (ResNet-18)	8.36	11.79	8.39	10.53
Ours	**6.79** ▼ 18.7%	**7.90** ▼ 33.0%	**7.38** ▼ 12.0%	**9.12** ▼ 13.4%
Ours (w/o adaptation)	7.16	8.63	7.89	9.26
Ours	**6.79** ▼ 5.2%	**7.90** ▼ 8.5%	**7.38** ▼ 6.5%	**9.12** ▼ 1.5%
Oracle	6.11	7.23	6.45	7.56

Table 2: Comparison results with baseline and oracle methods. We use either ETH-XGaze ($\mathcal{D}_E$) or Gaze360 ($\mathcal{D}_G$) as the labeled source data. The training split (903 persons) from GazeCapture ($\mathcal{D}_{GC}$) is used as the unlabeled person-specific data. We use either test split (109 persons) of GazeCapture ($\mathcal{D}_{GC}$) or MPIIGaze ($\mathcal{D}_M$) as the test data. Angular gaze error (°) is used as the evaluation metric.

5.1 Datasets and Setup

Since we propose a new problem setting, there have no existing datasets and experimental protocols for this problem. We re-purpose several existing gaze datasets for our problem, including ETH-XGaze [Zhang *et al.*, 2020], Gaze360 [Kellnhofer *et al.*, 2019], GazeCapture [Krafka *et al.*, 2016] and MPIIGaze [Zhang *et al.*, 2019]. The ETH-XGaze dataset ($\mathcal{D}_E$) contains 756,540 images without person IDs. The Gaze360 dataset ($\mathcal{D}_G$) has 84902 images without person IDs. The MPIIGaze dataset ($\mathcal{D}_M$) has 15 subjects, where each subject contains 3000 images. GazeCapture ($\mathcal{D}_{GC}$) is the largest in-the-wild dataset for gaze estimation. It contains data from over 1,450 people consisting of almost 2.5M frames. Since the original GazeCapture dataset only provides gaze labels on a 2D screen, we use the pre-processing pipeline [Park *et al.*, 2019] to attain 3D head pose from GazeCapture. We select 993 persons with over 400 samples per person as the unlabeled person-specific dataset for training. We select 109 persons with over 1000 samples per person are used for testing.

We perform experiments using the following different setups. We use either $\mathcal{D}_E$ or $\mathcal{D}_G$ as the labeled source data. The $\mathcal{D}_{GC}$ training split (993 persons) is used as the unlabeled person-specific data in all setups. We then test the model on either the $\mathcal{D}_{GC}$ test split (109 persons) or $\mathcal{D}_M$ (15 persons). These different setups will simulate real-world scenarios where data are combined from different sources.

5.2 Implementation Details

Network Architecture. We use a ResNet-based [He *et al.*, 2016] architecture as the main network B_ψ by removing the last fully-connected layer, then add a two-layer MLP with 256 hidden neurons and output dimension of 128 to the last block of ResNet. The main task outputs the 2-D dimension representing the yaw and pitch angles of the gaze. The auxiliary task (jigsaw puzzle) outputs a M-dimension vector representing the predicted permutation. During the training and testing, we only update the parameters for the last ResNet block and the MLP layer.

Training Details. We implement our method in PyTorch. All experiments are performed on a single RTX 3090 GPU. We use SGD for the meta-optimization with a fixed learning rate of $\beta = 10^{-4}$. The inner optimization is done using 3 gradient steps with an adaptation learning rate of $\alpha = 10^{-2}$. During meta-testing, we use the same fixed learning rate and

Category	Methods	Model	Target samples	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_{GC}$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_{GC}$
Without adaptation	ResNet-50 [Zhang *et al.*, 2020]	Agnostic	N/A	8.02	10.5	8.33	12.9
	Full-Face[Zhang *et al.*, 2017b]	Agnostic	N/A	12.35	14.52	11.13	16.12
	MANet[Wu *et al.*, 2022]	Agnostic	N/A	7.25	9.88	<u>7.96</u>	<u>9.86</u>
	Gaze360[Kellnhofer *et al.*, 2019]	Agnostic	N/A	<u>7.23</u>	<u>9.02</u>	11.36	15.86
	Baseline[Zhang *et al.*, 2020]	Agnostic	N/A	8.36	11.79	8.39	10.53
	Ours (w/o adaptation)	Agnostic	N/A	**7.16**	**8.63**	**7.89**	**9.26**
With adaptation	Gaze360[Kellnhofer *et al.*, 2019]	Dataset-specific	>100	10.23	9.82	7.48	<u>7.01</u>
	DAGEN[Guo *et al.*, 2020]	Dataset-specific	∼500	7.77	8.23	7.96	9.35
	GVBGD[Cui *et al.*, 2020]	Dataset-specific	∼1000	8.63	8.10	7.78	8.20
	UMA[Cai *et al.*, 2020]	Dataset-specific	∼100	9.96	9.42	8.37	7.54
	PnP-GA[Liu *et al.*, 2021]	Dataset-specific	∼100	<u>7.11</u>	<u>7.97</u>	**6.58**	**6.98**
	Ours	Person-specific	5	**6.79**	**7.90**	<u>7.38</u>	9.12

Table 3: Performance comparison with SOTA methods. Since our problem setting is new, we compare with existing supervised methods and domain adaptation methods. Our method achieves a competitive result. Note that the domain adaptation methods use more information than ours. In particular, they require access to the source domain and enough unlabeled data in the target domain during adaptation. They are also computationally expensive. Table 1 provides details for these settings. We report the angular error (°) as the evaluation metric.

3 gradient steps as meta-optimization to keep consistent with training. We set the weight of the distance between source and target data to $\gamma = 0.1$ (Eq. 4). The meta-batch size is set to $n = 10$ (Eq. 5) and each task consists of $K = 5$ face images. All images are cropped to the resolution of 224×224. For the auxiliary task, the face image is decomposed to 16 shuffled patches by a regular 4×4 grid, and we define $M = 31$ classes for the patch permutation prediction task.

5.3 Results and Comparison

Since this paper addresses a new problem in gaze estimation, there is no previous work that we can directly compare with. Nevertheless, we define several baseline methods for comparison. We also modify some state-of-the-art domain adaptation approaches for comparison.

Baseline: This method uses the same ResNet-18 [Zhang *et al.*, 2020] backbone as ours and is trained on the source data in a standard supervised manner. **Ours (w/o adaptation):** This is similar to our approach. The only difference is that it does not perform adaptation at test time. **Oracle:** We also consider an oracle method for comparison. This oracle is the same as ours during meta-training. But during meta-testing, it has access to the ground-truth labels of the images in the support set. It uses these labeled images (instead of unlabeled images in our case) to fine-tune the model with a supervised loss for adaptation. **PnP-GA:** [Liu *et al.*, 2021] This is a plug-and-play gaze estimation adaptation framework, which has access to 10 pre-trained models with about 100 target samples at test time. **DAGEN:** [Guo *et al.*, 2020] It is a state-of-the-art (SOTA) unsupervised domain adaptation gaze estimation model, which needs to access the 500 target samples for adaptation. **UMA:** [Cai *et al.*, 2020] It is an unsupervised domain adaptation using self-training to decrease domain shift, which needs around 100 target samples for adaptation. **GVBGD:** [Cui *et al.*, 2020] It is a SOTA unsupervised domain adaptation method for classification, which has access to 1000 target samples for adaptation.

Since some of above methods do not address gaze prediction (UMA [Cai *et al.*, 2020] and GVBGD [Cui *et al.*, 2020]), we retrain these models with gaze datasets. For a fair comparison, we use ResNet-18 (our backbone) to replace their original networks.

Comparison with Baselines. Table 2 shows the results of the proposed method and baseline methods. The baseline (ResNet-18) is trained on the source dataset (ETH-XGaze and Gaze360). "Ours" and "Ours (w/o adaptation" are trained on the labeled source dataset and unlabeled person-specific datasets (GazeCapture). "Oracle" is same as ours during meta-training, but it has access to the labels of the support set during meta-testing. So "Oracle" provides an upper bound for our approach. Our proposed method achieves significant improvement over the baseline. It also performs better than "Ours (w/o adaptation)" for in all cases.

Comparison with SOTA Methods. Table 3 shows the quantitative results of different state-of-the-art methods. We compare our method with four supervised gaze estimation methods, including ResNet-50 [Zhang *et al.*, 2020], Full-Face [Zhang *et al.*, 2017b], MANet [Wu *et al.*, 2022] and Gaze360 [Kellnhofer *et al.*, 2019]. The results are obtained from either the original papers or running official codes. We also compare with some domain adaptation methods, including Gaze360 [Kellnhofer *et al.*, 2019], DAGEN [Guo *et al.*, 2020], GVBGD [Cui *et al.*, 2020], UMA [Cai *et al.*, 2020] and PnP-GA [Liu *et al.*, 2021]. Note that domain adaption methods in Table 3 require more information that our method. All of them require enough unlabeled data (at least 100) in the target domain and access to the source data during adaptation. They also require heavy computation during the adaptation since they need to run gradient updates for many iterations. So Please refer to Table 1 for detailed comparisons of these settings. Nevertheless, the proposed model still achieves competitive performance compared with other SOTA methods.

6 Conclusion

We have proposed a new problem setting for efficient label-free user adaptation in gaze estimation. Previous work on user-adaptive gaze estimation requires labeled target data to fine-tune the model at test time. Besides, previous work requires large-scale training data containing gaze labels and person IDs. This is unrealistic in real-world scenarios. Our proposed new setting addresses these issues. We propose a novel meta-learning approach for this new problem. Our approach uses some labeled source data without person IDs and some unlabeled person-specific data for training. Our method uses a generalization bound from domain adaptation to define the loss function in meta-learning. This allows us to learn how to adapt to a new user with only a few unlabeled images even when we do not have training data annotated with both gaze labels and person IDs.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62372329), in part by Shanghai Rising Star Program (No.21QC1400900), in part by Tongji-Qomolo Autonomous Driving Commercial Vehicle Joint Lab Project, and in part by Xiaomi Young Talents Program.

References

[An *et al.*, 2021] Yuexuan An, Hui Xue, Xingyu Zhao, and Lu Zhang. Conditional self-supervised learning for few-shot classification. In *IJCAI*, 2021.

[Ben-David *et al.*, 2010] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

[Cai *et al.*, 2019] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. Learning disentangled semantic representation for domain adaptation. In *IJCAI*, 2019.

[Cai *et al.*, 2020] Minjie Cai, Feng Lu, and Yoichi Sato. Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In *IEEE CVPR*, 2020.

[Carlucci *et al.*, 2019] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *IEEE CVPR*, 2019.

[Chen *et al.*, 2020] Meixu Chen, Yize Jin, Todd Goodall, Xiangxu Yu, and Alan Conrad Bovik. Study of 3d virtual reality picture quality. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):89–102, 2020.

[Cheng *et al.*, 2020] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *AAAI*, 2020.

[Chi *et al.*, 2021] Zhixiang Chi, Yang Wang, Yuanhao Yu, and Jin Tang. Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning. In *IEEE CVPR*, 2021.

[Chi *et al.*, 2022] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. Metafscil: A meta-learning approach for few-shot class incremental learning. In *IEEE CVPR*, 2022.

[Chi *et al.*, 2024] Zhixiang Chi, Li Gu, Tao Zhong, Huan Liu, Yuanhao Yu, Konstantinos N Plataniotis, and Yang Wang. Adapting to distribution shift by visual domain prompt generation. In *ICLR*, 2024.

[Cui *et al.*, 2020] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *IEEE CVPR*, 2020.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[Funes Mora *et al.*, 2014] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. *Symposium on Eye Tracking Research and Applications*, 2014.

[Gidaris *et al.*, 2018] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

[Guo *et al.*, 2020] Zidong Guo, Zejian Yuan, Chong Zhang, Wanchao Chi, Yonggen Ling, and Shenghao Zhang. Domain adaptation gaze estimation by embedding with prediction consistency. In *ACCV*, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, 2016.

[He *et al.*, 2019] Junfeng He, Khoi Pham, Nachiappan Valliappan, Pingmei Xu, Chase Roberts, Dmitry Lagun, and Vidhya Navalpakkam. On-device few-shot personalization for real-time gaze estimation. In *IEEE ICCV Workshops*, 2019.

[Kellnhofer *et al.*, 2019] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *IEEE ICCV*, 2019.

[Konrad *et al.*, 2020] Robert Konrad, Anastasios Angelopoulos, and Gordon Wetzstein. Gaze-contingent ocular parallax rendering for virtual reality. *ACM Trans. Graph*, 39(2):1–12, 2020.

[Krafka *et al.*, 2016] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *IEEE CVPR*, 2016.

[Lee *et al.*, 2019] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *IEEE CVPR*, 2019.

[Liang *et al.*, 2020] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020.

[Liang *et al.*, 2022] Yudong Liang, Bin Wang, Wangmeng Zuo, Jiaying Liu, and Wenqi Ren. Self-supervised learning and adaptation for single image dehazing. In *IJCAI*, 2022.

[Lindén *et al.*, 2019] Erik Lindén, Jonas Sjöstrand, and Alexandre Proutiere. Learning to personalize in appearance-based gaze tracking. In *IEEE CVPR Workshops*, 2019.

[Liu *et al.*, 2021] Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *IEEE ICCV*, 2021.

[Liu *et al.*, 2024] Huan Liu, Julia Qi, Zhenhao Li, Mohammad Hassanpour, Yang Wang, Konstantinos Plataniotis, and Yuanhao Yu. Test-time personalization with meta prompt for gaze estimation. In *AAAI*, 2024.

[Long *et al.*, 2017] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, 2017.

[Long *et al.*, 2018] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

[Lu *et al.*, 2014] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Adaptive linear regression for appearance-based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(10):2033–2046, 2014.

[Mansour *et al.*, 2009] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *The 22nd Conference on Learning Theory*, 2009.

[Munkhdalai and Yu, 2017] Tsendsuren Munkhdalai and Hong Yu. Meta neworks. In *ICML*, 2017.

[Nichol *et al.*, 2018] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arxiv:1803.02999*, 2018.

[Park *et al.*, 2019] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *IEEE ICCV*, 2019.

[Peng *et al.*, 2024] Boyang Peng, Sanqing Qu, Yong Wu, Tianpei Zou, Lianghua He, Alois Knoll, Guang Chen, and changjun jiang. Map: Mask-pruning for source-free model intellectual property protection. In *IEEE CVPR*, 2024.

[Prange *et al.*, 2017] Alexander Prange, Michael Barz, and Daniel Sonntag. Speech-based medical decision support in vr using a deep neural network. In *IJCAI*, 2017.

[Qu *et al.*, 2022] Sanqing Qu, Guang Chen, Jing Zhang, Zhijun Li, Wei He, and Dacheng Tao. Bmd: A general class-balanced multicentric dynamic prototype strategy for source-free domain adaptation. In *ECCV*, 2022.

[Shen *et al.*, 2021] Chengchao Shen, Xinchao Wang, Youtan Yin, Jie Song, Sihui Luo, and Mingli Song. Progressive network grafting for few-shot knowledge distillation. In *AAAI*, 2021.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[Sun *et al.*, 2020] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020.

[Sung *et al.*, 2018] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *IEEE CVPR*, 2018.

[Tang *et al.*, 2021] Xuwen Tang, Zhu Teng, Baopeng Zhang, and Jianping Fan. Self-supervised network evolution for few-shot classification. In *IJCAI*, 2021.

[Vinyals *et al.*, 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[Wang *et al.*, 2020] Wenguan Wang, Jianbing Shen, Xingping Dong, Ali Borji, and Ruigang Yang. Inferring salient objects from human fixations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(8):1913–1927, 2020.

[Wu *et al.*, 2022] Yong Wu, Gongyang Li, Zhi Liu, Mengke Huang, and Yang Wang. Gaze estimation via modulation-based adaptive network with auxiliary self-learning. *IEEE Trans. Circuit Syst. Video Technol.*, 32:5510–5520, 2022.

[Wu *et al.*, 2023] Yong Wu, Shekhor Chanda, Mehrdad Hosseinzadeh, Zhi Liu, and Yang Wang. Few-shot learning of compact models via task-specific meta distillation. In *IEEE WACV*, 2023.

[Wu *et al.*, 2024] Yanan Wu, Zhixiang Chi, Yang Wang, Konstantinos N. Plataniotis, and Songhe Feng. Test-time domain adaptation by learning domain-aware batch normalization. In *AAAI*, 2024.

[Xu *et al.*, 2020] Renjun Xu, Pelen Liu, Yin Zhang, Fang Cai, Jindong Wang, Shuoying Liang, Heting Ying, and Jianwei Yin. Joint partial optimal transport for open set domain adaptation. In *IJCAI*, 2020.

[Yu *et al.*, 2019] Yu Yu, Gang Liu, and Jean-Marc Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *IEEE CVPR*, 2019.

[Zhang *et al.*, 2017a] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Everyday eye contact detection using unsupervised gaze target discovery. In *ACM Symposium on User Interface Software and Technology*, 2017.

[Zhang *et al.*, 2017b] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *IEEE CVPR Workshops*, 2017.

[Zhang *et al.*, 2019] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(1):162–175, 2019.

[Zhang *et al.*, 2020] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *ECCV*, 2020.

[Zhong *et al.*, 2022] Tao Zhong, Zhixiang Chi, Li Gu, Yang Wang, Yuanhao Yu, and Jin Tang. Meta-dmoe: Adapting to domain shift by meta-distillation from mixture-of-experts. In *NeurIPS*, 2022.