MoEGaze: A Mixture of Experts Approach for Generalizable Gaze Estimation

Zheng Liu, Feng Lu*

State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University
{liu_zheng, lufeng}@buaa.edu.cn

Abstract

Existing gaze estimation models often struggle to generalize to unseen users, primarily due to significant variations in individual appearance. Empirical observations reveal that performance improves when the visual appearance of test subjects closely resembles that of training subjects. Motivated by this, we propose a generalizable gaze estimation framework MoEGaze based on the Mixture of Experts (MoE) architecture. During training, the model extracts appearance features from facial images and uses them to route samples to specialized gaze expert networks, each tailored to a specific subset of appearances. Rather than directly predicting gaze, each expert outputs intermediate gaze features, which are dynamically aggregated according to the input appearance and then mapped to gaze prediction. This dynamic routing design enables the model to effectively adapt to users with diverse appearances, while also facilitating easier training on sub-datasets with smaller appearance variations. Extensive experiments demonstrate that our method achieves superior cross-domain performance compared to existing approaches, with an average improvement of 27.6% across four cross-domain metrics over the baseline. Furthermore, MoEGaze surpasses baselines trained on the full dataset while requiring only 10% of the training data.

Introduction

Human gaze carries rich attentional information (Frischen, Bayliss, and Tipper 2007) and reflects cognitive states (Calder et al. 2002), serving as a crucial behavioral cue. Recent years have witnessed growing applications of gaze estimation in intelligent cockpits (Doshi and Trivedi 2009), mixed reality (Blattgerste, Renner, and Pfeiffer 2018; Piumsomboon et al. 2017), and human-computer interaction (Piumsomboon et al. 2017; Cai et al. 2025b). While traditional near-infrared-based solutions face scalability challenges due to hardware costs (Cheng et al. 2021; Guestrin and Eizenman 2006), appearance-based methods that directly regress gaze direction from webcam images (Zhang et al. 2015; Hisadome et al. 2023; Liu, Wang, and Lu 2024) demonstrate promising potential for widespread deployment.

However, appearance-based approaches require extensive training data. Existing datasets (Zhang et al. 2020; Kellnhofer et al. 2019; Krafka et al. 2016), limited by collection costs, cannot cover all user demographics and environmental conditions, leading to significant performance degradation in new domains. This challenge stems from distributional shifts across datasets in terms of user characteristics and environmental factors. Current solutions mainly fall into two categories: 1) Unsupervised domain adaptation methods (Kellnhofer et al. 2019; Guo et al. 2021; Bao et al. 2022) that adapt models using target domain images, though such images are often unavailable; and 2) Domain generalization approaches that enhance transferability through feature purification (Cheng, Bao, and Lu 2022), contrastive learning (Wang et al. 2022a), or vision-language priors (Yin et al. 2024), typically at the cost of within-domain performance.

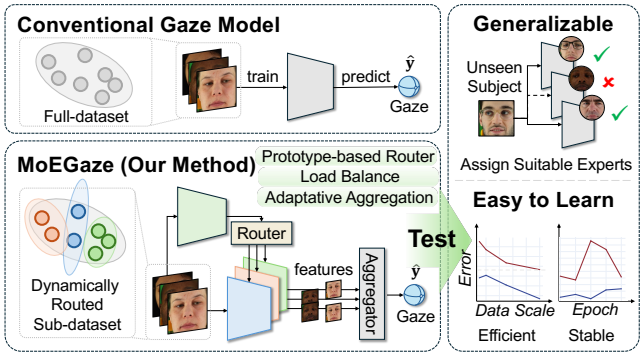

Figure 1: The simplified overall architecture of the proposed Mixture-of-Experts-based gaze estimation algorithm (MoEGaze). Unlike conventional gaze models that learn directly from the entire dataset, MoEGaze employs dynamic routing to partition the dataset into multiple subsets, with an expert model trained on each subset.

These methods overlook a critical aspect: optimal utilization of source domain data. Our key insight is that models trained with target-user-similar data outperform those trained on generic datasets. This motivates our core methodology: partitioning training data into specialized subsets for training expert models, with automatic selection of the most distribution-aligned expert during inference.

The recent success of Mixture-of-Experts (MoE) in large language models (Fedus, Zoph, and Shazeer 2022; Lepikhin

*Corresponding author

et al. 2020) validates its effectiveness. MoE employs gating networks to: 1) cluster input data, and 2) dynamically activate experts. Its strength lies in enabling experts to specialize in particular data distributions, achieving optimal sub-task performance - a principle perfectly aligned with our insight.

Building on this, we propose MoEGaze (illustrated in Fig. 1), which innovatively adapts the MoE framework to gaze estimation. Our solution: 1) dynamically partitions training data by appearance features, 2) trains specialized experts on each subset, and 3) intelligently matches the optimal expert during inference. This design not only reduces learning difficulty but also significantly improves generalization to unseen users. The main contributions of this work are summarized as follows:

- We identified a strong correlation between cross-subject gaze estimation error and the appearance feature divergence between training and testing data, demonstrating that models trained with appearance-similar data consistently achieve superior performance.
- Building on this insight, we propose MoEGaze, an innovative and generalizable gaze estimation framework capable of simultaneously extracting both appearance features and gaze-related features from images. Our framework dynamically partitions the training data to develop appearance-specialized experts and employs an adaptive routing mechanism to select the most suitable expert during inference.
- We design an prototype-based routing mechanism that innovatively combines historical assignment with current appearance features for expert selection, outperforming conventional routers in both performance and interpretability.
- Extensive experiments demonstrate that our method achieves state-of-the-art performance, with improvements of 12.5%, 30.1%, 21.6%, and 38.6% on $\mathcal{D}_E \rightarrow \mathcal{D}_M$, $\mathcal{D}_E \rightarrow \mathcal{D}_D$, $\mathcal{D}_G \rightarrow \mathcal{D}_M$, and $\mathcal{D}_G \rightarrow \mathcal{D}_D$ tasks compared to baseline, respectively. Notably, our approach demonstrates remarkable data efficiency, surpassing the performance of the baseline trained on the full dataset while using only 10% of the training samples. This highlights the inherent optimization efficiency of the proposed method.

Related Works

Generalizable Gaze Estimation

Gaze estimation has attracted increasing research attention in recent years. While early approaches relied on handcrafted features, modern deep learning-based methods can automatically learn effective gaze regression features from large-scale data (Zhang et al. 2015, 2017; Cheng and Lu 2021). Advanced computer vision techniques have been progressively incorporated into gaze estimation research: Yu *et al.* (Yu and Koltun 2016) employed dilated convolutions for improved gaze prediction; Cheng *et al.* (Cheng and Lu 2021) proposed a hybrid CNN-Transformer architecture; and Yin *et al.* (Yin et al. 2024) achieved state-of-the-art performance by leveraging pretrained vision-language models.

However, collecting large-scale gaze estimation datasets remains prohibitively expensive, requiring numerous participants and sophisticated systems (Krafka et al. 2016; Kellnhofer et al. 2019; Zhang et al. 2020). This motivates the development of algorithms that can learn more generalizable features from smaller datasets. Domain generalization has emerged as a prominent solution: Cheng *et al.* (Cheng, Bao, and Lu 2022) purified gaze-related features via adversarial learning; Xu *et al.* (Xu, Wang, and Lu 2023) systematically investigated how data augmentation and adversarial learning perturb gaze-irrelevant features; Wang *et al.* (Wang et al. 2022b) proposed a contrastive learning framework for regression tasks. Nevertheless, current approaches suffer from two limitations: 1) training difficulty due to full-dataset dependency, and 2) compromised within-domain performance. Thus, developing more trainable gaze estimation algorithms with strong both in- and cross-domain performance remains an important research direction.

Mixture-of-Experts Architecture

The Mixture of Experts (MoE) paradigm operates on the principle of partitioning training data into subsets, enabling each expert model to specialize in its corresponding data distribution (Vats et al. 2024). This concept traces back to early machine learning studies (Jordan and Jacobs 1994; Jacobs et al. 1991). With the advent of neural networks, Eigen *et al.* (Eigen, Ranzato, and Sutskever 2013) pioneered the integration of MoE into deep learning, proposing architectures with multiple routers and experts. A breakthrough came from Shazeer *et al.* (Shazeer et al. 2017), who dramatically improved the training efficiency of massive neural networks through sparse gating mechanisms. Recently, MoE has achieved remarkable success not only in large language models but also in computer vision (Ahmed, Baig, and Torresani 2016; Liao et al. 2025) and knowledge transfer (Zhong et al. 2023). To our knowledge, no prior work has applied the MoE framework to gaze estimation, while this integration holds significant potential for addressing gaze estimation's generalization challenges.

Motivation

Facial Appearance and Gaze Estimation

Extensive research has confirmed that facial appearance variations substantially impair gaze estimation accuracy. Current solutions typically employ unsupervised domain adaptation or supervised fine-tuning with target-user data. Given that target-user-specific models outperform those trained on generic datasets, this raises a fundamental research question: *Does an optimal balance exist between target and generic data that maximizes performance for target users?* To investigate, we train individualized models using each subject's data, evaluating on two challenging subgroups: African-descent individuals (underrepresented with skin-tone-related bias) and glasses-wearers (where frames occlude facial regions). We compare three model categories: 1) *self-models* (trained on target individual's data), 2) *in-group models* (trained on subgroup-specific data), and 3) *random models* (trained on randomly sampled data). Fig. 2

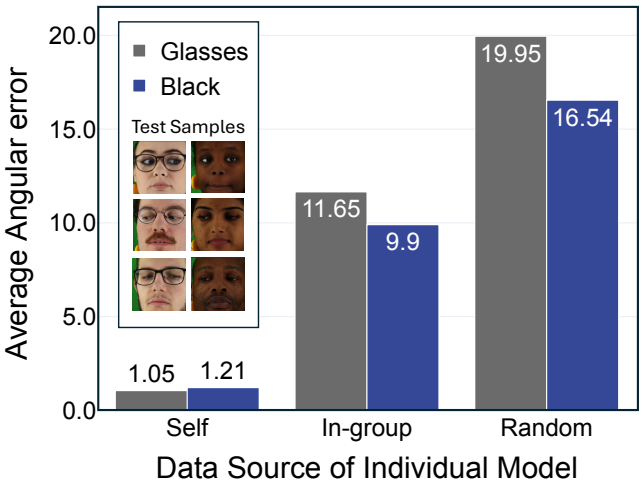

Figure 2: Average test results of two appearance sets evaluated three types of individually trained models (trained on single-subject data). Here, **Self** indicates models trained with the test subject's own data, **In-group** indicates models trained on data from individuals sharing the same appearance group as the test subject, and **Random** indicates models trained on data from randomly selected subjects.

Method	Test Group	
	Glasses	Black
Baseline	7.41	5.97
Full-data Fine-tuning	7.64 ▲ 3.1%	5.84 ▼ 2.2%
In-group Fine-tuning	**7.19** ▼ **3.0%**	**5.76** ▼ **3.5%**

Table 1: Performance comparison of naive fine-tuning on the entire dataset versus fine-tuning using data from the target appearance group.

demonstrates that in-group models (appearance-matched) consistently surpass random models, proving that while appearance discrepancy degrades performance, appearance consistency leads to better results.

Training on Similar-Appearance Subjects

Building upon these experimental observations, we investigate whether demographic-aware training (i.e., using data from users with similar appearance characteristics to test subjects) can enhance gaze estimation accuracy. Using a baseline model pretrained on a subset of the ETH-XGaze dataset, we evaluate performance on two held-out test cases: (1) black participants and (2) glasses-wearing participants. We then fine-tune the baseline separately with three datasets: (a) four black users, (b) four glasses-wearing users, and (c) the complete training set for comparison. As shown in Table 1, appearance-specific (in-group) fine-tuning achieves optimal performance gains. These findings demonstrate the importance of appearance-matched training data for optimal model performance.

Mixture-of-Experts-based Gaze Estimation

Based on the discussion above, we conclude that splitting the dataset into subsets according to image appearance and training specialized sub-models can lead to more generalizable gaze estimation. Inspired by the success of Mixture-of-Experts (MoE) architectures in the field of large language models (Cai et al. 2025a; Fedus, Zoph, and Shazeer 2022; Lepikhin et al. 2020), we propose MoEGaze, an MoE-based gaze estimation algorithm designed to enable seamless training of appearance-specific gaze feature extractors (i.e., gaze experts). Compared to conventional gaze estimation models, MoEGaze is more specialized in handling diverse appearances and can generalize better to unseen subjects by dynamically allocating appropriate gaze experts. During training, MoEGaze actively partitions the dataset into appearance-similar subsets and trains experts on these subsets. The gaze features generated by these experts are then dynamically aggregated using a decoder-style aggregator. The overall architecture is illustrated in Fig. 3.

MoEGaze: a MoE-based Appearance-Adaptive Gaze Estimation Framework

Task Definition

The objective of gaze estimation generalization is to improve model performance on unseen users beyond the training distribution. To achieve this, we are typically given a source domain dataset $\mathcal{D}_s = \{(\mathbf{x}_s, \mathbf{y}_s)\}_{i=1}^{N}$ for training a gaze estimation model $\mathbf{G}(\mathbf{x} \mid \Theta)$. During training, only data from $\mathcal{D}_s$ is available, and the process optimizes the following objective: $\arg\min_{\Theta} \mathcal{L}(\mathbf{y}_s, \mathbf{G}(\mathbf{x}_s \mid \Theta))$, where $\mathcal{L}$ denotes the loss function. Algorithm performance is evaluated on multiple target datasets $\{\mathcal{D}_t^k\}_{k=1}^{K} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ by computing the angular error E between predicted and ground truth gaze vectors, usually defined as $E(\mathbf{y}_i, \hat{\mathbf{y}}_i) = \frac{1}{N}\sum_{i=1}^{N}\langle \mathbf{y}_i, \hat{\mathbf{y}}_i \rangle$. An effective domain generalization algorithm should maintain strong performance across all target datasets, demonstrating robust generalization to unseen users and scenarios.

Overall Architecture

To effectively categorize individuals by their appearance types and train specialized models for each type, we propose a gaze estimation network based on the Mixture-of-Experts (MoE) architecture, inspired by its successful applications in large language models. As shown in Fig. 3, the proposed model consists of three key components: a set of gaze expert networks, an appearance expert network, and an adaptive aggregation module.

MoGE To learn the gaze representation of different types of appearance separately, we proposed **M**ixture **o**f **G**aze **E**xperts (MoGE) network. As shown in Fig. 3, the MoGE module typically comprises N visual backbone networks serving as gaze feature extractors $\mathcal{E}$ (*i.e.* gaze expert). These experts are independently trained on N mutually exclusive subsets $\mathcal{D}_s^i$ of the source domain dataset $\mathcal{D}_s$ to enhance inter-expert diversity. As illustrated in Fig. 3, the MoGE architecture takes an image as input and outputs N sets of gaze

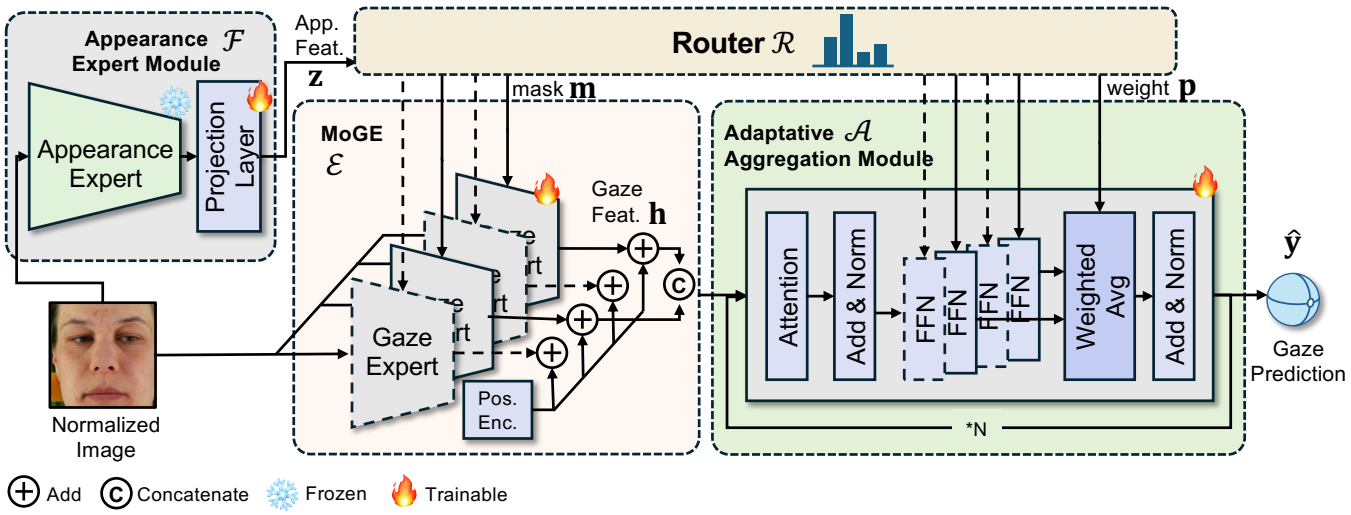

Figure 3: Overview of the proposed MoEGaze, comprising four components: the Appearance Expert Module, Mixture of Gaze Experts (MoGE), Adaptive Aggregation Module, and a Router. Given a normalized image, the Appearance Expert extracts appearance features, which the Router uses to select relevant gaze experts. Activated experts extract gaze features, which are positionally encoded and adaptively aggregated based on expert scores to produce the final gaze prediction.

features with their mapped gaze directions:

$$\mathbf{g}_i, \mathbf{h}_i = \mathcal{E}_i(\mathbf{x}_s) \tag{1}$$

During inference, a sparse activation mechanism is employed—only a selected subset of K experts $\mathcal{E}'_i$ determined by the routing module is activated, thereby substantially reducing the computational overhead.

Appearance Expert Module We introduce a dedicated appearance expert module to extract facial appearance features for both expert model routing and adaptive feature aggregation guidance. The module architecture builds upon a standard Vision Transformer (ViT) backbone (Dosovitskiy et al. 2021), initialized with pre-trained weights from (Zheng et al. 2022). These weights were optimized through language-image contrastive learning, enabling the capture of rich semantic information from facial images - particularly advantageous for semantic-driven appearance classification in our framework. The complete appearance expert module, illustrated in Fig. 3, comprises: 1) a frozen ViT feature extractor to preserve semantic knowledge, and 2) a trainable MLP head that projects the semantic features into our task-specific appearance representation space. The appearance feature extraction process is formally defined as:

$$\mathbf{z} = \mathcal{F}(\mathbf{x}_s) \tag{2}$$

Adaptive Aggregation Module. Given multiple trained gaze experts, fusing their extracted gaze features $\mathbf{h}_i$ proves more effective than simply selecting the optimal expert's output. We therefore propose an adaptive aggregation module that integrates selected experts' features via self-attention and maps them to gaze direction vectors. However, vanilla self-attention suffers from two limitations: 1) identical expert groups should maintain consistent fusion strategies, and 2) the combinatorial explosion of possible expert selections makes predefined fusion modules infeasible. Inspired by Switch Transformer (Fedus, Zoph, and Shazeer 2022), we incorporate sparse-activated feedforward layers and dynamic routing into the attention module, enabling appearance-adaptive feature fusion.

A critical challenge arises from directly fusing expert outputs: the selected experts and their ordering may vary across inputs. We address this by augmenting each expert's features $\mathbf{h}_i$ with sinusoidal positional encodings PE_i, yielding $\mathbf{h}'_i = \mathbf{h}_i + PE_i$ to preserve feature correspondence.

The sparse-activated feedforward network contains parallel MLPs corresponding to each expert. When the routing mask $\mathbf{m}_j = 1$, the associated MLP activates, enabling dynamic fusion:

$$\mathbf{h}'_i = \frac{1}{K} \sum_{j:\mathbf{m}_j=1} \mathbf{p}_{ij} \mathrm{FFN}_j(\mathbf{h}'_i) \tag{3}$$

The final gaze direction prediction combines averaged fused features through a linear layer:

$$\hat{\mathbf{y}} = \mathbf{W} \cdot \frac{1}{K} \sum_{i=1}^{K} \mathbf{h}'_i \tag{4}$$

where $\mathbf{W}$ denotes the learnable weight matrix.

Prototype-based Router

To enable appearance-based expert assignment, we propose a Prototype-based Router. This design is motivated by the key observation in prior section: models trained predominantly on specific facial appearance types demonstrate superior performance on similar appearances. This suggests expert selection should incorporate historical assignment patterns. Existing MoE routing methods (Fedus, Zoph, and Shazeer 2022; Lepikhin et al. 2020) neither support appearance-based selection nor maintain historical information, necessitating our novel router.

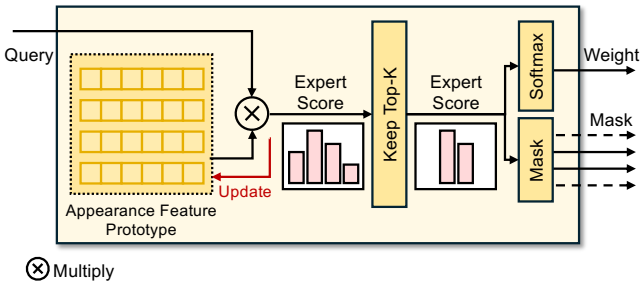

Figure 4: Detail of Prototype-based Router.

The core innovation is a dynamically maintained expert appearance prototype matrix $\mathbf{Z}_{proto} \in \mathbb{R}^{N \times D}$, where each prototype represents the centroid of appearances best handled by corresponding experts, computed from historical assignments. For input appearance feature $\mathbf{z}$, expert scores are obtained via cosine similarity with prototypes:

$$\mathbf{p} = \mathbf{Z}_{proto} \cdot \frac{\mathbf{z}}{\|\mathbf{z}\|} \tag{5}$$

We preserve the top-K scores ($K = 3$ empirically) followed by Softmax normalization:

$$\mathbf{p} = \text{Softmax}(\text{KeepTopK}(\mathbf{p}, K)) \tag{6}$$

where KeepTopK operates as:

$$\text{KeepTopK}_i(\mathbf{p}, K) = \begin{cases} \mathbf{p}_i, & \text{if } \mathbf{p}_i \in \text{top}K(\mathbf{p}) \\ -\infty, & \text{otherwise} \end{cases} \tag{7}$$

The expert selection mask $\mathbf{m}$ is generated as:

$$\mathbf{m}_i = \begin{cases} 1, & \text{if } \mathbf{p}_i \in \text{top}K(\mathbf{p}) \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Finally, selected prototypes are updated via momentum ($\alpha = 0.01$):

$$\mathbf{Z}^i_{proto} \leftarrow \mathbf{Z}^i_{proto} + \mathbf{m}_i \cdot \alpha \left(\frac{\mathbf{z}}{\|\mathbf{z}\|} - \mathbf{Z}^i_{proto} \right) \tag{9}$$

Alignment During the training phase, we optimize the routing module to dynamically assign input images to their most suitable expert models, which essentially aligns the routing score distribution with expert performance. Since different experts exhibit varying prediction accuracy for the same input, we quantify their performance using the angular prediction error $\mathbf{e}_i$:

$$\mathbf{e}_i = \langle \mathbf{W} \cdot \mathbf{h}_i, \mathbf{y}_i \rangle \tag{10}$$

To establish alignment between routing scores and expert performance, we propose a cross-entropy based alignment loss. Specifically, we first normalize both the expert angular errors $\mathbf{e}$ and routing scores $\mathbf{p}$ using Softmax functions, then minimize their distribution discrepancy:

$$\mathcal{L}_{ali} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{K} \mathbf{e}_{ij} \log \frac{\exp(\mathbf{p}_{ij})}{\sum_{k=1}^{K} \exp(\mathbf{p}_{ik})} \tag{11}$$

Load Balance Unconstrained expert assignment by appearance features may cause load imbalance, compromising model capacity and underutilizing rarely-selected experts (Fedus, Zoph, and Shazeer 2022). We address this via a dual-constraint mechanism combining: (1) balance loss for equitable expert selection frequency, and (2) entropy regularization to preserve decision discriminability, preventing degenerate random routing.

1) Balance Loss: For batch size B and N experts, let $\mathbf{p}_{ij}$ denote the probability of sample $\mathbf{x}_j$ selecting expert $\mathcal{E}_i$ ($\sum_i \mathbf{p}_{ij} = 1$). When choosing K experts, the expected selection count per expert is $\mathbf{s}_i = \sum_j \mathbf{p}_{ij}$. The balanced state requires:

$$\mathbf{s}_i \approx \frac{BK}{N}, \quad \forall i \tag{12}$$

Normalizing $\mathbf{q}_i = \mathbf{s}_i/(BK)$, we measure deviation from uniform distribution $\mathbf{u}_i = 1/N$ via KL divergence:

$$\mathcal{L}_{bal} = D_{KL}(\mathbf{q} \parallel \mathbf{u}) = \log N + \sum_i \mathbf{q}_i \log \mathbf{q}_i \tag{13}$$

2) Entropy Regularization: Sole balance loss may drive $\mathbf{p}_{ij}$ toward $1/N$, nullifying expert selection. Inspired by RL (Schulman et al. 2017; Mnih et al. 2016), we employ entropy regularization to boost decision confidence:

$$\mathcal{L}_{ent} = -\frac{1}{B} \sum_j \sum_i \mathbf{p}_{ij} \ln \mathbf{p}_{ij} \tag{14}$$

Total Loss

For the gaze direction vector obtained from aggregated features, we employ angular error as the supervision signal. This error, combined with individual experts' prediction errors $\mathbf{e}$, forms the gaze loss $\mathcal{L}_{gaze}$:

$$\mathcal{L}_{gaze} = \frac{1}{K+1} \left(\sum_{i=1}^{K} \mathbf{e}_i + \langle \hat{\mathbf{y}}, \mathbf{y} \rangle \right) \tag{15}$$

The final objective function is a weighted sum of all loss components:

$$\mathcal{L}_{total} = \mathcal{L}_{gaze} + \lambda_1 \mathcal{L}_{ali} + \lambda_2 \mathcal{L}_{bal} + \lambda_3 \mathcal{L}_{ent} \tag{16}$$

Here, λ_1, λ_2 and λ_3 are empirically set to $\lambda_1 = 1$, $\lambda_2 = 1$ and $\lambda_3 = 10$.

Experiments

Datasets

To validate the cross-domain generalization capability of our MoEGaze method, we conduct experiments on four commonly-used datasets: ETH-XGaze (Zhang et al. 2020) ($\mathcal{D}_E$), Gaze360 (Kellnhofer et al. 2019) ($\mathcal{D}_G$), MPIIFaceGaze (Zhang et al. 2017) ($\mathcal{D}_M$), and EyeDiap (Funes Mora, Monay, and Odobez 2014) ($\mathcal{D}_D$). Based on their inherent characteristics, we select $\mathcal{D}_E$ and $\mathcal{D}_G$ with wider head pose variations as training sets, and establish four cross-domain evaluation protocols: $\mathcal{D}_E \rightarrow \mathcal{D}_M$, $\mathcal{D}_E \rightarrow \mathcal{D}_D$, $\mathcal{D}_G \rightarrow \mathcal{D}_M$, and $\mathcal{D}_G \rightarrow \mathcal{D}_D$. All datasets are preprocessed following (Zhang et al. 2017) with standard operations including image rotation, scale normalization, and corresponding label transformations.

Comparison Methods

Our framework is evaluated against representative methods from three primary paradigms in gaze estimation: (1) *General-purpose vision backbones*: They can extract general visual features suitable for most computer vision tasks, demonstrating consistent and robust performance in gaze estimation. We evaluate four models: ResNet18 (He et al. 2016), ResNet50 (He et al. 2016), ViT-H (Dosovitskiy et al. 2021), and Swin Transformer (Liu et al. 2021). (2) *Dedicated gaze estimation networks*: They are designed for within-domain cross-subject scenarios but with limited cross-domain generalization. We assess three representative models: Full-Face (Zhang et al. 2017), Dilated-Net (Chen and Shi 2019), and CA-NET (Cheng et al. 2020). (3) *Cross-domain generalization methods*: They often leverage specialized learning strategies (e.g., feature purification, adversarial attack, etc.) to extract domain-invariant gaze features. We comprehensively compare four state-of-the-art (SOTA) approaches: PureGaze (Cheng, Bao, and Lu 2022), CDG (Wang et al. 2022a), Gaze-Consistent (Xu, Wang, and Lu 2023), and CLIP-Gaze (Yin et al. 2024).

Implementation Details

Please refer to supplementary materials for more details.

Domain Generalization Ability

To validate the effectiveness of our proposed gaze estimation network MoEGaze, we compare it with SOTA methods under four standard cross-domain benchmarks (Table 2). While general-purpose vision backbones like Swin-Transformer extract transferable features and achieve a promising 7.61° mean error, MoEGaze consistently yields better performance, indicating stronger cross-domain generalization. Compared to the ResNet-50 baseline, our method improves accuracy by nearly 3°, enhancing adaptability in real-world scenarios.

Domain-specific gaze models often struggle with generalization, leading to suboptimal performance in our evaluations. In contrast, MoEGaze significantly outperforms these methods. When compared with SOTA cross-domain gaze estimation approaches, MoEGaze matches the best results of SOTA on $\mathcal{D}_E \rightarrow \mathcal{D}_M$ and $\mathcal{D}_G \rightarrow \mathcal{D}_D$, while surpassing them on the other two settings. Overall, it achieves the best average cross-domain generalization performance, highlighting its superior robustness across diverse domains.

Within Domain Performance

Thanks to MoEGaze's dynamic appearance-based expert assignment, our method excels in both cross-domain and within-domain gaze estimation. To validate within domain performance, we conduct comprehensive comparisons with CDG, ResNet-50 baseline, and PureGaze under within-domain settings (see Table 3). Notably, all compared methods adopt ResNet-50 as their backbone architecture. Experimental results reveal that existing cross-domain methods suffer performance drops—especially PureGaze, whose accuracy nearly matches its cross-domain results. In contrast, MoEGaze achieves a 0.4° improvement, validating the effectiveness of our appearance-based expert assignment strategy.

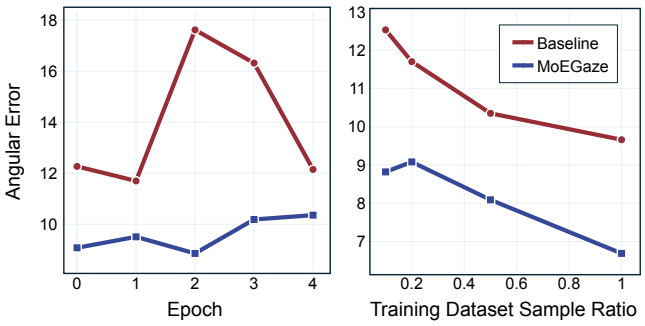

Figure 5: Results of ResNet-18-based MoEGaze trained on a sampled dataset, compared with the baseline (ResNet-18). The plots show cross-domain errors from $\mathcal{D}_E \rightarrow \mathcal{D}_D$. Left: training process using 20% of the dataset. Right: model performance across different dataset sizes.

Learned on Sampled Dataset

The design of MoEGaze enables each expert to specialize in similar-appearance subsets, allowing the model to achieve competitive performance even with limited training data. To verify this, we sample 10%, 20%, and 50% of each user's training data from the ETH dataset, and train corresponding ResNet-18-based MoEGaze. For comparison, a ResNet-18 baseline pretrained on the same data as expert pretraining is also trained on different scales of dataset. All models are evaluated on the $\mathcal{D}_E \rightarrow \mathcal{D}_D$ task (see Fig. 5).

Remarkably, MoEGaze trained with merely 10% data outperforms the baseline using full dataset. Furthermore, cross-domain errors across different epochs reveal that while both approaches exhibit overfitting tendencies, the baseline demonstrates pronounced training instability while MoEGaze maintains stable convergence. We attribute this phenomenon to reduced gaze-irrelevant variations in appearance-similar subsets, which facilitates more robust learning of gaze-related features.

Ablation Study

Loss Function We conduct an ablation study on loss functions using a progressive setup across four cross-domain tasks, starting with the basic gaze loss $\mathcal{L}_{gaze}$. As shown in Table 4, using only $\mathcal{L}_{gaze}$ yields competitive results. Adding $\mathcal{L}_{bal}$ notably improves performance on $\mathcal{D}_G \rightarrow \mathcal{D}_D$, while causing a slight degradation on $\mathcal{D}_G \rightarrow \mathcal{D}_M$. After adding $\mathcal{L}_{ent}$, the degradation was eliminated, and the performance on $\mathcal{D}_E \rightarrow \mathcal{D}_M$ is further improved. The full loss combination achieves the best results on all tasks, with improvements of 7.3%, 4.9%, 4.3%, and 11.9%, averaging a 7.3% gain over the baseline.

Expert Configuration We evaluate combinations of expert count $N \in \{4, 8\}$ and activation number K on $\mathcal{D}_E \rightarrow \mathcal{D}_M$ and $\mathcal{D}_E \rightarrow \mathcal{D}_D$ (see Table 5). The best average error occurs at $N = 4$, $K = 3$, challenging the assumption that more experts always improve performance. We suggest that excessive N reduces data routed to per expert, harming training under limited data. When $N = 8$, increasing K consistently

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$	Avg.
ResNet-18 (He et al. 2016)	8.35	9.66	7.58	9.01	8.65
ResNet-50 (He et al. 2016)	7.59	8.70	8.75	11.83	9.21
ViT-H (Dosovitskiy et al. 2021)	7.68	8.58	8.24	7.79	8.07
Swin-S (Liu et al. 2021)	7.73	7.91	7.21	7.61	7.61
Full-Face (Zhang et al. 2017)	12.35	30.15	11.13	14.42	17.01
Dilated-Net (Chen and Shi 2019)	-	-	18.45	23.88	21.16
CA-NET (Cheng et al. 2020)	-	-	27.13	31.41	29.27
PureGaze (Cheng, Bao, and Lu 2022)	7.08	7.48	9.28	9.32	8.29
CDG (Wang et al. 2022a)	6.73	7.95	7.03	7.27	7.25
CLIP-Gaze (Yin et al. 2024)	**6.41**	7.51	6.89	**7.06**	6.97
Gaze-Consistent (Xu, Wang, and Lu 2023)	6.50	7.44	7.55	9.03	7.63
MoEGaze (Ours)	6.64	**6.08**	**6.86**	7.26	**6.71**

Table 2: Domain generalization performance comparison with state-of-the-art methods

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_E$
ResNet-50 (baseline) (He et al. 2016)	4.5
PureGaze (Cheng, Bao, and Lu 2022)	7.6
CDG (Wang et al. 2022a)	4.6
MoEGaze (ours)	**4.1**

Table 3: Within-domain performance comparison between the proposed method, baseline, and other state-of-the-art domain generalization gaze estimation methods.

Loss	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$	Avg.
$\mathcal{L}_{gaze}$	7.16	6.39	7.17	8.24	7.24
$+\mathcal{L}_{bal}$	7.16	6.25	7.54	7.66	7.15
$+\mathcal{L}_{ent}$	6.87	6.20	7.16	7.60	6.95
$+\mathcal{L}_{ali}$	**6.64**	**6.08**	**6.86**	**7.26**	**6.71**

Table 4: Ablation study on the proposed loss functions.

improves results; however, this trend vanishes for $N = 4$. We hypothesize that: (1) Higher K mitigates data sparsity at $N = 8$, while (2) too large K at $N = 4$ weakens dynamic routing, reducing MoEGaze to simple self-attention aggregation.

Hyperparameters in Aggregation Module We systematically study two key hyperparameters in the adaptive aggregation module: the number of aggregation layers L and attention heads H. Experiments vary $L \in \{1, 4, 6, 12\}$ with fixed $H = 8$, and $H \in \{1, 4, 8, 16\}$ with fixed $L = 6$ (see Table 6). On $\mathcal{D}_E \rightarrow \mathcal{D}_M$, the error decreases then increases with L, reaching 6.64° at $L = 6$; on $\mathcal{D}_E \rightarrow \mathcal{D}_D$, the metrics show no significant variation. Increasing H shows a U-shaped trend on both $\mathcal{D}_E \rightarrow \mathcal{D}_M$ and $\mathcal{D}_E \rightarrow \mathcal{D}_D$. The best average performance is achieved at $L = 6, H = 8$. Compared to expert count N and selection number K, L and H have a smaller effect on accuracy.

Experts	Activate Num	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	Average
N=4	K=1	8.62	7.23	7.92
	K=2	6.65	6.12	6.38
	K=3	**6.64**	**6.08**	**6.36**
	K=4	6.58	6.55	6.56
N=8	K=2	7.77	8.41	8.09
	K=4	7.65	6.46	7.05
	K=6	7.57	6.78	7.17
	K=8	7.35	6.65	7.00

Table 5: Ablation on expert configuration.

		$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	Average
Layer L	L=1	7.17	6.07	6.62
	L=4	7.01	6.05	6.53
	L=6	**6.64**	6.08	**6.36**
	L=12	7.37	**6.02**	6.69
Heads H	H=1	7.39	6.38	6.88
	H=4	7.34	**5.91**	6.62
	H=8	**6.64**	6.08	**6.36**
	H=16	6.91	6.21	6.56

Table 6: Hyperparameter ablation of the Adaptive Aggregation Module.

Conclusion

This paper presents MoEGaze, a novel gaze estimation algorithm that dynamically splits training data into similar-appearance subsets and trains specialized expert models on each partition. Comprehensive evaluations demonstrate that MoEGaze achieves state-of-the-art performance, with consistent improvements across all four cross-domain benchmarks and within-domain settings. Remarkably, our framework exhibits exceptional data efficiency - models trained with merely 10% of the dataset surpass baselines trained on complete data, underscoring the effectiveness of MoEGaze.

Acknowledgments

This work was supported by Beijing Natural Science Foundation (L242019).

References

Ahmed, K.; Baig, M. H.; and Torresani, L. 2016. Network of Experts for Large-Scale Image Categorization. In *European Conference on Computer Vision*, 516–532. Springer.

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.

Blattgerste, J.; Renner, P.; and Pfeiffer, T. 2018. Advantages of eye-gaze over head-gaze-based selection in virtual and augmented reality under varying field of views. In *Proceedings of the workshop on communication by gaze interaction*, 1–9.

Cai, W.; Jiang, J.; Wang, F.; Tang, J.; Kim, S.; and Huang, J. 2025a. A Survey on Mixture of Experts in Large Language Models. *IEEE Transactions on Knowledge and Data Engineering*, 1–20.

Cai, Z.; Hong, J.; Wang, Z.; and Lu, F. 2025b. GazeSwipe: Enhancing Mobile Touchscreen Reachability through Seamless Gaze and Finger-Swipe Integration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–14.

Calder, A. J.; Lawrence, A. D.; Keane, J.; Scott, S. K.; Owen, A. M.; Christoffels, I.; and Young, A. W. 2002. Reading the mind from eye gaze. *Neuropsychologia*, 40(8): 1129–1138.

Chen, Z.; and Shi, B. E. 2019. Appearance-Based Gaze Estimation Using Dilated-Convolutions. arXiv:1903.07296.

Cheng, Y.; Bao, Y.; and Lu, F. 2022. PureGaze: Purifying Gaze Feature for Generalizable Gaze Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1): 436–443.

Cheng, Y.; Huang, S.; Wang, F.; Qian, C.; and Lu, F. 2020. A Coarse-to-Fine Adaptive Network for Appearance-Based Gaze Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07): 10623–10630.

Cheng, Y.; and Lu, F. 2021. Gaze Estimation Using Transformer. arXiv:2105.14424.

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

Doshi, A.; and Trivedi, M. M. 2009. On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes. *IEEE Transactions on Intelligent Transportation Systems*, 10(3): 453–462.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Eigen, D.; Ranzato, M.; and Sutskever, I. 2013. Learning Factored Representations in a Deep Mixture of Experts. *arXiv preprint arXiv:1312.4314*.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

Frischen, A.; Bayliss, A. P.; and Tipper, S. P. 2007. Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin*, 133(4): 694.

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. Safety Harbor Florida: ACM. ISBN 978-1-4503-2751-0.

Guestrin, E.; and Eizenman, M. 2006. General Theory of Remote Gaze Estimation Using the Pupil Center and Corneal Reflections. *IEEE Transactions on Biomedical Engineering*, 53(6): 1124–1133.

Guo, Z.; Yuan, Z.; Zhang, C.; Chi, W.; Ling, Y.; and Zhang, S. 2021. Domain Adaptation Gaze Estimation by Embedding with Prediction Consistency. In Ishikawa, H.; Liu, C.-L.; Pajdla, T.; and Shi, J., eds., *Computer Vision – ACCV 2020*, volume 12626, 292–307. Cham: Springer International Publishing. ISBN 978-3-030-69540-8 978-3-030-69541-5.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. Las Vegas, NV, USA: IEEE. ISBN 978-1-4673-8851-1.

Hisadome, Y.; Wu, T.; Qin, J.; and Sugano, Y. 2023. Rotation-Constrained Cross-View Feature Fusion for Multi-View Appearance-based Gaze Estimation. arXiv:2305.12704.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive Mixtures of Local Experts. *Neural computation*, 3(1): 79–87.

Jordan, M. I.; and Jacobs, R. A. 1994. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural computation*, 6(2): 181–214.

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.

Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; and Torralba, A. 2016. Eye Tracking for Everyone. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2176–2184. Las Vegas, NV, USA: IEEE. ISBN 978-1-4673-8851-1.

Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2020. Gshard: Scaling Giant Models with Conditional Computation and Automatic Sharding. *arXiv preprint arXiv:2006.16668*.

Liao, M.; Dong, H. B.; Wang, X.; Ubul, K.; Yan, Z.; and Shao, Y. 2025. GM-MoE: Low-light Enhancement

with Gated-Mechanism Mixture-of-Experts. *arXiv preprint arXiv:2503.07417*.

Liu, R.; Wang, H.; and Lu, F. 2024. From Gaze Jitter to Domain Adaptation: Generalizing Gaze Estimation by Manipulating High-Frequency Components. *International Journal of Computer Vision*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002. Montreal, QC, Canada: IEEE. ISBN 978-1-66542-812-5.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 1928–1937. PmLR.

Piumsomboon, T.; Lee, G.; Lindeman, R. W.; and Billinghurst, M. 2017. Exploring natural eye-gaze-based interaction for immersive virtual reality. In *2017 IEEE symposium on 3D user interfaces (3DUI)*, 36–39. IEEE.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *arXiv preprint arXiv:1701.06538*.

Vats, A.; Raja, R.; Jain, V.; and Chadha, A. 2024. The Evolution of Mixture of Experts: A Survey from Basics to Breakthroughs.

Wang, Y.; Jiang, Y.; Li, J.; Ni, B.; Dai, W.; Li, C.; Xiong, H.; and Li, T. 2022a. Contrastive Regression for Domain Adaptation on Gaze Estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19354–19363. New Orleans, LA, USA: IEEE. ISBN 978-1-66546-946-3.

Wang, Y.; Jiang, Y.; Li, J.; Ni, B.; Dai, W.; Li, C.; Xiong, H.; and Li, T. 2022b. Contrastive Regression for Domain Adaptation on Gaze Estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19354–19363. New Orleans, LA, USA: IEEE. ISBN 978-1-66546-946-3.

Xu, M.; Wang, H.; and Lu, F. 2023. Learning a Generalized Gaze Estimator from Gaze-Consistent Feature. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3): 3027–3035.

Yin, P.; Zeng, G.; Wang, J.; and Xie, D. 2024. CLIP-Gaze: Towards General Gaze Estimation via Visual-Linguistic Model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7): 6729–6737.

Yu, F.; and Koltun, V. 2016. Multi-Scale Context Aggregation by Dilated Convolutions. arXiv:1511.07122.

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 Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, volume 12350, 365–381. Cham: Springer International Publishing. ISBN 978-3-030-58557-0 978-3-030-58558-7.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2015. Appearance-Based Gaze Estimation in the Wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4511–4520. Boston, MA, USA: IEEE. ISBN 978-1-4673-6964-0.

Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2299–2308. Honolulu, HI, USA: IEEE. ISBN 978-1-5386-0733-6.

Zheng, Y.; Yang, H.; Zhang, T.; Bao, J.; Chen, D.; Huang, Y.; Yuan, L.; Chen, D.; Zeng, M.; and Wen, F. 2022. General Facial Representation Learning in a Visual-Linguistic Manner. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18676–18688. New Orleans, LA, USA: IEEE. ISBN 978-1-66546-946-3.

Zhong, T.; Chi, Z.; Gu, L.; Wang, Y.; Yu, Y.; and Tang, J. 2023. Meta-DMoE: Adapting to Domain Shift by Meta-Distillation from Mixture-of-Experts. arXiv:2210.03885.