FEW-SHOT GAZE ESTIMATION WITH MODEL OFFSET PREDICTORS

Jiawei Ma[⋆†] *Xu Zhang*[†] *Yue Wu*[†] *Varsha Hedau*[†] *Shih-Fu Chang*[⋆†]

[⋆] Columbia University
[†] Amazon Alexa AI

ABSTRACT

Due to the variance of optical properties across different people, the performance of a person-agnostic gaze estimation model may not generalize well on a specific person. Though one may achieve better performance by training a person-specific model, it typically requires a large number of samples which is not available in real-life scenarios. Hence, few-shot gaze estimation method is preferred for the small number of samples from a target person. However, the key question is how to close the performance gap between a "few-shot" model and the "many-shot" model. In this paper, we propose to learn a person-specific offset predictor which outputs the difference between the person-agnostic model and the many-shot person-specific model with as few as one training sample. We adapt the knowledge to a new person by using the average of meta-learned offset predictors parameters as the initialization of the new offset predictor. Experiments show that the proposed few-shot person-specific model is not only closer to the corresponding many-shot person-specific model but also has better accuracy than the SOTA few-shot gaze estimation methods in multiple gaze datasets.

***Index Terms*—** Few-shot learning, gaze estimation, initialization, consistency

1. INTRODUCTION

Gaze estimation aims at detecting the direction of the human gaze from a face image or video. Recently, with the rapid development of deep learning technologies [1], the appearance-based gaze estimation (estimate gaze from face images) methods [2, 3] have surpassed the conventional model-based (construct a 3D eye model to estimate the gaze) methods [4]. A number of large-scale gaze datasets [5, 6, 3] with a wide collection of subjects have been proposed to train a robust gaze estimation CNN model. A gaze estimation model trained on the dataset without differentiating subjects is therefore a person-agnostic model. However, due to the variance across persons, such as eye-ball shape and face appearance, the precision of person-agnostic gaze models is still limited [7].

A model which can explore more optical characteristics for one specific person generally has better accuracy and is required in many real-world applications. However, to ensure a good model, collecting thousands of samples [7] from each target person to fulfill the training is required, which can be very expensive and not suitable to real-life applications. To reduce the data collection burden, the idea of meta-learning [8, 9] with few-shot learning becomes appealing.

Meta-learning contains two major phases in the context of gaze estimation [9, 10], meta-training and meta-testing. During the meta-training phase, a model is learned by training across all person-specific tasks from all person in the training set, which is then assumed to be generalizable to new persons (a.k.a. meta-model). In the meta-testing phase, one derives a few-shot model from the meta-model for a target person with a few training samples and is expected to achieve accurate prediction. The state-of-the-art few-shot gaze estimation method FAZE [10] applies the Model-Agnostic Meta-Learning (MAML) [9] approach. It sup-samples the full training set to simulate few-shot tasks (*i.e.*, each task has a limited number of training samples and a reserved test set) to learn the meta-model. Then, the meta model is used as the initialization of a gaze estimation model where an adapted model to a new person is derived by finetuning with a few samples.

In our research, we noticed that the parameters of the final few-shot gaze estimation model has large discrepancy to the parameters of the corresponding many-shot person-specific model (see Tab. 3), where a many-shot model is a concept opposed to a few-shot model and is trained with many more samples (e.g., 500). A deeper look reveals the reason that FAZE samples few-shot training task in meta-training but doesn't explicitly make the few-shot model close to the many-shot model, which results in deficient accuracy of FAZE. We argue that, learning a few-shot model which is closed to its corresponding many-shot model can improve the performance of the few-shot model.

To achieve that, for each person, we propose to learn a person-specific offset predictor to estimate the offset of model parameters between the many-shot person-specific model and the person-agnostic model in meta-learning. We further utilize a consistency loss to enhance that the offset predictor outputs similar offsets with any training samples from the target person, which helps provide a consistent prediction result. During meta-testing, inspired by [11], we leverage the knowledge in meta-training by using the average of the offset pre-

This work was done when Jiawei was an intern at Amazon Alexa AI.

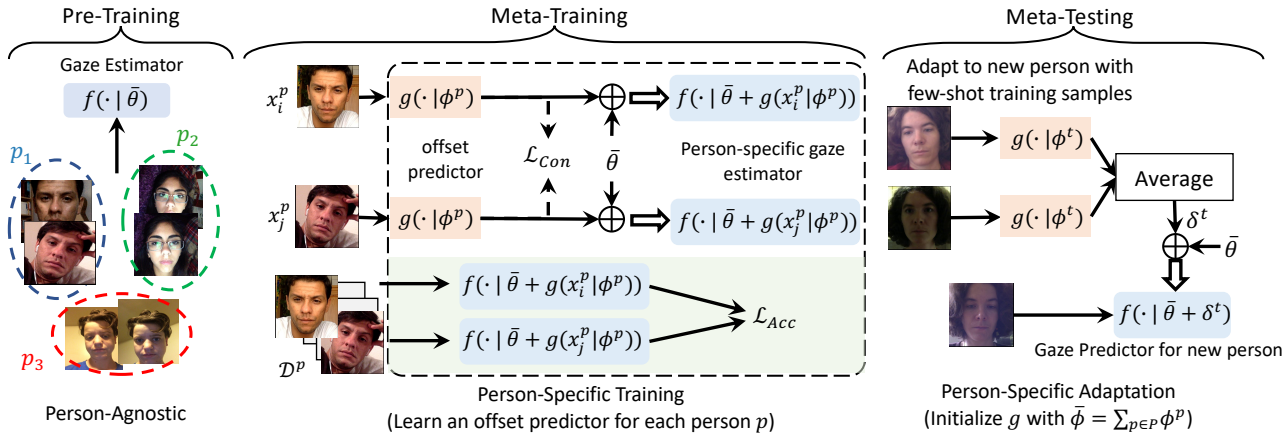

Fig. 1: Pre-Training: we first learn a person-agnostic model $f(\cdot|\bar{\theta})$ across all samples. Meta-Training: for each person p, an offset predictor $g(\cdot|\phi^p)$ is trained to generate the offset of the model parameter w.r.t. $\bar{\theta}$. Meta-Testing: given a new person, we initialize the predictor with the average of all person-specific parameters and then finetune it with a few training samples.

dictors as the initialization. Experiment shows, the few-shot model learned by the proposed method is not only closer to the many-shot model but also achieves better performance.

The paper makes the following contributions:

- We propose a new few-shot gaze estimation strategy to make the few-shot model close to the many-shot model.
- We propose to learn a offset predictor with a consistency loss to predict the offset of the model parameters between the person-specific many-shot model and the person-agnostic model.
- The proposed algorithm achieves significant performance gain compared to the SOTA few-shot gaze estimation methods in multiple gaze datasets.

2. APPROACH

2.1. Formulation

Given a training dataset with M people, the training samples with regard to a person p can be represented as $\mathcal{D}^p = \{(\mathbf{x}_i^p, y_i^p)\},\ i \in [1, N^p]$, where $\mathbf{x}_i^p$ is an image, y_i^p is its corresponding gaze label (a 3-dimension unit vector indicating the gaze direction), and N^p is the number of training samples for person p. All the people forms a set $\mathcal{P}, |\mathcal{P}| = M$. The whole training set can be represented as $\mathcal{D} = \cup_{p\in\mathcal{P}}\mathcal{D}^p$. Given a new test person t (who never appears in $\mathcal{P}$), with K training samples $\mathcal{D}^t = \{(\mathbf{x}_i^t, y_i^t)\},\ i \in [1, K]$, the goal of few-shot gaze estimation is to train a person-specific gaze estimation model $f(\cdot|\theta^t)$ for t by leveraging information in $\mathcal{D}$ and $\mathcal{D}^t$, where θ^t is the parameter for the model.

2.2. Overview

The flow graph of the proposed algorithm is shown in Fig. 1. We first pre-train a person-agnostic gaze estimation model $f(\cdot|\bar{\theta})$ on all samples in $\mathcal{D}$. It generally has a good generalization ability but a deficient accuracy for each specific person.

During meta-training, for each person $p \in \mathcal{P}$, we start training a person-specific model with all available training samples $\mathcal{D}^p$ from the person-agnostic one. Instead of directly fine-tuning from $f(\cdot|\bar{\theta})$, we propose to learn an offset predictor $g(\cdot|\phi^p)$, which is parameterized by ϕ^p. It takes an arbitrary image $\mathbf{x}_i^p \in \mathcal{D}^p$ as input and its output serves as the model parameter offset between the person-agnostic model ($\bar{\theta}$) and the person-specific model (θ^p), *i.e.*, $\theta^p = \bar{\theta} + g(\mathbf{x}_i^p|\phi^p)$. As the few-shot samples can be diverse and unrepresentative, the predictor is thus be trained to output consistent offsets. In this way, the predictor provides more stable prediction than the conventional finetuning process.

During meta-testing, we take the average of the offset predictor parameters from different people in the training set as network initialization. Since the averaged parameter has the minimum average squared Frobenius norm to all the person-specific offset predictors [11], it can be easily adapted to a person-specific offset predictor with only few training samples from a new test person.

2.3. Meta-Training

Starting with the person-agnostic model, for each person p, we train a offset predictor $g(\cdot|\phi^p)$ with parameter ϕ^p to dynamically predict the offsets w.r.t. $\bar{\theta}$ to obtain the many-shot person-specific model θ^p. Therefore, we are capable to regularize more on the offsets comparing with directly updating the offsets value through gradient back-propagation. The predicted offsets are expected to 1) build a person-specific gaze estimator achieving high accuracy for a target person (*Accuracy*), while 2) the offsets predicted from various training samples should always be close to each other (*Consistency*).

Given a person $p \in \mathcal{P}$ and its dataset $\mathcal{D}^p$, regarding for

the *Accuracy* goal, the predicted gaze should be as close as the ground-truth label. Then, we use ℓ_2 loss function, *i.e.*,

$$L_{\text{Acc}} = \frac{1}{(N^p)^2} \sum_{i=1}^{N^p} \sum_{j=1}^{N^p} \parallel y_i^p - f\big(\mathbf{x}_i^p | \bar{\theta} + g(\mathbf{x}_j^p | \phi^p)\big) \parallel_2^2 .$$

With the predicted offset $g(\mathbf{x}_j^p|\phi^p)$, the gaze prediction for sample $\mathbf{x}_i^p$ is $f(\mathbf{x}_i^p|\bar{\theta} + g(\mathbf{x}_j^p|\phi^p))$ where the gaze estimator is parameterized by $\bar{\theta} + g(\mathbf{x}_j^p|\phi^p)$. Note that the conventional many-shot fine-tuning training is a special case of the above formulation where $g(\mathbf{x}_j^p|\phi^p)$ always gives a constant offset.

To make the offset prediction consistent, we propose to minimize the variance [12] of different predicted offsets from the same person. The consistency loss can be denoted as

$$L_{\text{Con}} = \frac{1}{N^p} \sum_{i=1}^{N^p} \parallel g(\mathbf{x}_i^p|\phi^p) - \frac{1}{N^p} \sum_{j=1}^{N^p} g(\mathbf{x}_j^p|\phi^p) \parallel_2^2 .$$

The final loss function to train the predictor is

$$L = L_{\text{Acc}} + \lambda L_{\text{Con}}, \tag{1}$$

where λ is a trade-off parameter. In practice, due to computational resource limitation, we calculate Eq. 1 within each batch. Since there are a total of M people in the dataset, we have M person-specific offset predictors with the parameters $\{\phi^p\}_{p\in\mathcal{P}}$ after previous steps. Also, we apply data augmentation to the training sample such that even if there is only 1 training image, we can still calculate the consistency loss.

By averaging all parameters from the M predictors, *i.e.*, $\bar{\phi} = \frac{1}{M}\sum_{p\in\mathcal{P}} \phi^p$, we use $\bar{\phi}$ as a meta offset predictor model for initialization before meta-testing. Our assumption is, $\bar{\phi}$ has the minimum average squared Frobenius norm [13] to the parameters of all the trained person-specific offset predictors. Therefore, given a new person with limited number of samples, the offset predictor should be easily adapted while keep both the accuracy and the consistency properties.

2.4. Meta-Testing

As mentioned above, given a new person t with K training samples $\mathcal{D}^t$, we initialize the offset predictor $g(\cdot)$ with $\bar{\phi}$ and tune the predictor with $\mathcal{D}^t$ using Eq. 1. Assuming the fine-tuned model is denoted as $g(\cdot|\phi^t)$, the predicted offset for sample $\mathbf{x}_i^t$ is then $g(\mathbf{x}_i^t|\phi^t)$. Finally, the average of offsets δ^t is considered as the offset for the specific person, *i.e.*,

$$\delta^t = \frac{1}{K} \sum\nolimits_{i=1}^{K} g(\mathbf{x}_i^t|\phi^t). \tag{2}$$

The offset is then to be added to the person-agnostic model $\bar{\theta}$ to build a person-specific gaze estimator $f(\cdot|\bar{\theta} + \delta^t)$ for evaluation. Note, instead of lively predicting each offset for each input testing sample, we choose to freeze the offset using the few-shot training sample. We noticed that it makes the final prediction more stable.

Table 1: Performance comparison with SOTA methods. Number shows the angular error (↓) in degree.

	Method \ Shot K	0	1	5	10	15
GazeCapt.	Appe*[16]	**3.47**	–	–	5.50	4.86
	Diff*[17]	–	4.68	3.80	3.64	3.57
	FAZE [10]	3.49	3.66	3.08	2.99	2.99
	Ours	3.50	**2.88**	**2.70**	**2.65**	**2.63**
MPIIGaze	Appe*[16]	**5.01**	–	6.07	4.07	3.68
	Diff*[17]	–	4.63	3.70	3.52	3.44
	FAZE [10]	5.23	3.91	3.24	3.14	3.12
	Ours	5.08	**3.33**	**2.91**	**2.84**	**2.81**

* Results of Appe [16] and Diff [17] are reported in [10].

Table 2: Ablation study. Number shows the angular error (↓) in degree.

Method	GazeCapture			MPIIGaze		
	K=1	K=5	K=10	K=1	K=5	K=10
No-OP-M	3.35	2.97	2.81	3.72	3.21	3.03
No-OP-FT	3.11	2.89	2.85	3.70	3.14	3.06
No-CL	3.07	2.77	2.74	3.48	3.16	3.07
Rand-Int	3.73	3.03	2.99	5.20	3.48	3.42
Ours	**2.88**	**2.70**	**2.65**	**3.33**	**2.91**	**2.84**

3. EXPERIMENT

3.1. Dataset and Implementation Details

We use **GazeCapture [6]** and **MPIIGaze [3]** for evaluation. Following the splits in Faze [10], we use 795 subjects in GazeCapture ($\approx 1.3M$ image samples) for training. During evaluation, 109 subjects in GazeCapture and 15 subjects in MPIIGaze are used for meta-testing. For each person, we use the last 500 samples (ordered by filenames) as test set and sample the few-shot training set from the remaining images. All images in these two datasets are captured in real life. However, since the two datasets are gathered separately, there is domain shift (*e.g.*camera and environments) between each other. Also, since the network is only trained on GazeCapture, the evaluation on MPIIGaze also shows the adaptation ability of the few-shot approaches to a new domain.

For fair comparison, we extract features for images using the backbone in [10]. Then, we set a two-layer MLP, with SeLU activation function [14], as the gaze estimator $f(\cdot)$. Another two-layer MLP is set as offset predictor $g(\cdot)$ to predict offsets for $f(\cdot)$. The offset prediction is only trained for the gaze estimator not the feature backbone. During training, we use Adam optimizer [15] and set λ as 1.0. For each person, we randomly sample the training samples and train the few-shot model for 10 times and report the average performance[1].

[1]Implementation and visualization results can be found in `https://github.com/Phoenix-V/fsl-gaze-offset`

3.2. Result

Compared with SOTA. We compare our method with multiple baselines including the state-of-the-art few-shot gaze estimation algorithm (FAZE [10]) in Table 1. For each dataset, we vary the number of training samples $K \in \{0, 1, 5, 10, 15\}$, where $K = 0$ is the evaluation by the model without any adaptation (the person-agnostic model for our method). From the table, the proposed algorithm is a clear winner among all the algorithms. With one training sample ($K = 1$), it outperforms FAZE by reducing the angular error for 0.78 degree (a relative 21.3%) in the GazeCapture dataset and 0.58 degree (a relative 14.8%) in the MPIIGaze dataset.

Although the model performance keeps increasing with larger K, it begins to saturate after $K = 5$. Since collecting person-specific training sample is expensive, finding a good trade-off K is important. Finally, the angular error in the MPIIGaze dataset is larger compared to that in GazeCapture, which indicates the impact of domain shift, since the model is only trained with the GazeCapture dataset.

Ablation Study. As shown in Tab. 2, we provide ablation study results for each component to analyze the performance gain. No-OP-M and No-OP-FT mean training the person-specific model *without* using the Offset Predictor. Instead, MAML and Fine-Tuning are used respectively. No-CL means model trained with the offset predictor but *without* Consistency Loss. Rand-Int means using the Random weights, instead of the averaged predictor parameters $\bar{\phi}$, to initialize the offset predictor before meta-testing.

No-OP-M and No-OP-FT both outperform FAZE and differ from FAZE on the gaze estimator initialization. FAZE directly learns a meta-model for the gaze estimator by training from the sub-sampled few-shot tasks, while the others start with a person-agnostic model $\bar{\theta}$. As such, even though directly testing the meta-model in FAZE [10] achieves similar performance as testing with $\bar{\theta}$ ($K = 0$ in Tab. 1), the person-agnostic model still serves as a better initialization to facilitate the adaptation before finetuning. Meanwhile, as Rand-Int works the worst, we find that the offset predictor initialization $\bar{\phi}$ in our method does bring the information learned on train set in meta-training step to the meta-testing step. Finally, the full model outperforms all baselines and shows the effectiveness of our proposed ideas.

Relation between Few-Shot and Many-Shot Models. For each person in meta-testing, we trained a person-specific many-shot model with all of its training samples (> 500), which is then compared with ten 5-shot models of the same person. The l_2-norm and CKA [18] (mean and standard deviation) between the model parameters are used to measure the similarity between the many-shot and few-shot models. As shown in Tab. 3, we can clearly see that our few-shot models are much closer to the many-shot models compared to the few-shot models in FAZE. The consistency loss further reduces the discrepancy.

Table 3: The Relation between the many-shot model and few-shot models. Numbers are average values on two datasets.

Method(5-shot)	FAZE[10]	No-CL	Ours
l_2-norm (↓)	$3.428_{\pm 0.006}$	$1.078_{\pm 0.016}$	$\mathbf{0.984}_{\pm 0.008}$
CKA (↑)	$0.963_{\pm 0.005}$	$0.997_{\pm 0.001}$	$\mathbf{0.998}_{\pm 0.001}$

Table 4: Angular error under different λ values.

λ	0	0.1	0.5	1.0	2.0	10.0
Error (↓)	2.84	2.86	2.82	**2.81**	2.86	3.41

Table 5: Angular error (↓) with different offset predictors.

GazeCapture	FC	2-layer MLP	3-layer MLP
$K = 1$	3.02	2.88	**2.85**
$K = 5$	2.79	2.70	**2.69**

Ablation study on λ. Table. 4 shows the mean angular error (averaged on two datasets) by varying the trade-off parameter λ from 0 to 10. When λ is less than 10, the performance is relatively stable but the minimum error is achieved when λ is 1. When λ is high, enforcing large weight to minimize diversity may make the model overfit and hurt the performance.

Offset Predictor $g(\cdot)$: Table. 5 shows the mean angular errors on GazeCapture with $g(\cdot)$ of different layers. With more layers, the computation workload increases and $g(\cdot)$ has stronger capability. However, we only see a marginal gain between the 2-layer MLP and the 3-layer MLP. Besides, a more complex model has a higher risk to overfit, especially with limited number of training samples. Thus, we choose the 2-layer MLP to balance the performance and the complexity.

Complexity Analysis. Compared to FAZE [10], due to the additional offset prediction step, the proposed approach adds additional computation to the meta-learning steps. However, since the offset predictor is only a 2-layer MLP, the impact on overall training process is very limited. Finally, we freeze the obtained offsets. To test a new sample, the gaze estimation complexity is exactly the same as a regular one.

4. CONCLUSION

Few-shot learning is an efficient way to learn a person-specific model. In this paper, we first discuss the discrepancy between SOTA few-shot model and the many-shot model. Then, for each person, we propose to fill the gap by learning a specific predictor to estimate the offset of gaze estimator parameters between a person agnostic model and the person-specific many-shot model. The offset predictor is initialized with the average of all person-specific predictor parameters learned on training set and then adapted to a new person. The proposed method significantly reduces the gap between the many-shot and few-shot models and improves the gaze prediction accuracy.

5. REFERENCES

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT press, 2016.

[2] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu, "Appearance-based gaze estimation with deep learning: A review and benchmark," *arXiv preprint arXiv:2104.12668*, 2021.

[3] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling, "Appearance-based gaze estimation in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4511–4520.

[4] Anuradha Kar and Peter Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16495–16519, 2017.

[5] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[6] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2176–2184.

[7] Elias Daniel Guestrin and Moshe Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE Transactions on biomedical engineering*, vol. 53, no. 6, pp. 1124–1133, 2006.

[8] Jake Snell, Kevin Swersky, and Richard S Zemel, "Prototypical networks for few-shot learning," *arXiv preprint arXiv:1703.05175*, 2017.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[10] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz, "Few-shot adaptive gaze estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9368–9377.

[11] Vipul Gupta, Santiago Akle Serrano, and Dennis DeCoste, "Stochastic weight averaging in parallel: Large-batch training that generalizes well," *arXiv preprint arXiv:2001.02312*, 2020.

[12] Alexander McFarlane Mood, "Introduction to the theory of statistics.," 1950.

[13] Roger A Horn and Charles R Johnson, *Matrix analysis*, Cambridge university press, 2012.

[14] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter, "Self-normalizing neural networks," in *Proceedings of the 31st international conference on neural information processing systems*, 2017, pp. 972–981.

[15] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] Xucong Zhang, Yusuke Sugano, and Andreas Bulling, "Evaluation of appearance-based methods and implications for gaze-based applications," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–13.

[17] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez, "A differential approach for gaze estimation with calibration.," in *BMVC*, 2018, vol. 2, p. 6.

[18] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton, "Similarity of neural network representations revisited," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3519–3529.