2025 IEEE International Conference on Image Processing (ICIP) | 979-8-3315-2379-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICIP55913.2025.11084699

DFT GAZE: DISTILLED AND FINE-TUNED GAZE ESTIMATION FOR PERSONALIZATION ON TINY DEVICES

He-Yen Hsieh[1], *Ziyun Li*[2], *Sai Qian Zhang*[2,3], *Wei-Te Mark Ting*[1], *Kao-Den Chang*[1],
Barbara De Salvo[2], *Chiao Liu*[2], *H. T. Kung*[1]

[1]Harvard University [2]Reality Labs Research, Meta [3]New York University

ABSTRACT

Real-time personalized gaze estimation on AR/VR devices requires both accuracy and efficiency, especially when adapting to individual users with limited personal data. This task is challenging due to low-latency requirements, the presence of dataset biases from dominant gaze directions, and risk of catastrophic forgetting during adaptation. We present Distilled and Fine-Tuned (DFT) Gaze, a lightweight model for personalized gaze estimation. Distilled from a larger teacher model, DFT Gaze reduces model size while retaining essential visual features through knowledge distillation, without relying on gaze-specific supervision. During fine-tuning, it integrates gaze-specific supervision with Adapters, reaching 281K parameters for efficient adaptation and online updates on edge devices. To mitigate dataset biases and reduce catastrophic forgetting, we introduce a clustering-based sampling that balances gaze distribution for better generalization and improves adaptation to individual gaze patterns, even with only 5 personal images. DFT Gaze outperforms state-of-the-art methods on the MPIIFaceGaze dataset for personalized gaze estimation. Despite having the smallest model size at 281K parameters, it maintains low gaze errors across other datasets, including MPIIGaze, OpenEDS2020, and AEA. At 10× smaller than its teacher model, DFT Gaze achieves fast inference, a low parameter count, and effective adaptation, making it well-suited for real-time applications in resource-constrained environments.

Index Terms— Real-time personalized gaze estimation, Self-supervised model distillation, Imbalanced gaze distribution, Catastrophic forgetting

1. INTRODUCTION

Gaze estimation determines the direction of a person's gaze from eye or face images, enabling applications in augmented and virtual reality (AR/VR) [1, 2], mental health assessment [3, 4], and human-computer interaction [5]. Real-time applications require gaze models that are both efficient and accurate. Recent trends [6–8] in gaze estimation commonly use large models trained on extensive datasets, which deliver strong performance but come with high computational costs, making them impractical for real-time and edge device deployment. Generalized gaze estimation aims to provide robust predictions across diverse users but often overfits to dominant gaze directions, struggling with rare variations. Personalized gaze estimation [6, 7] fine-tunes models for individual users to account for variations in eye shapes and movements, improving accuracy but requiring substantial user data and risking overfitting when trained on limited samples.

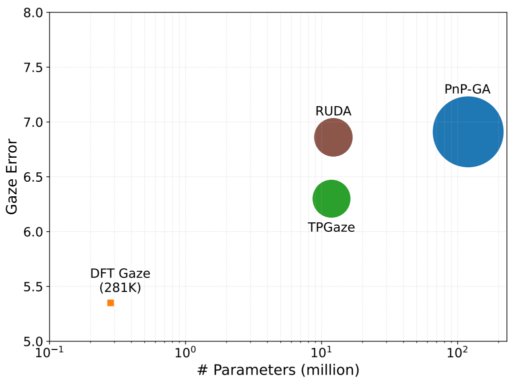

Fig. 1. Gaze error versus model size. Our model achieves lower gaze error than other approaches on MPIIFaceGaze, despite having only 281K parameters.

To address these challenges, we introduce DFT Gaze, an ultra-compact gaze estimation model that optimizes efficiency and accuracy through both structured knowledge distillation (KD) and clustering-based fine-tuning. A major limitation of gaze datasets for pretraining is their lack of diversity, as they are often collected under controlled conditions with limited lighting variations and background complexity. Models trained solely on gaze data often struggle with real-world variability due to limited diversity in lighting, background complexity, and environmental conditions. Therefore, we use knowledge distillation (KD) with a masked autoencoder to transfer knowledge from a large teacher model to a compact student model. The student learns broad visual representations without direct gaze supervision, improving its robustness to different lighting conditions, backgrounds, and environments in real-world gaze estimation.

Fine-tuning on gaze datasets presents additional challenges. In generalized gaze estimation, gaze directions are often imbalanced, with common directions dominating fine-tuning. This creates gradient bias and reduces the model's ability to generalize to less frequent gaze angles. In personalized gaze estimation, fine-tuning on a small number of user-specific samples increases the risk of overfitting, making it harder for the model to adapt beyond the user's typical gaze patterns. To address these issues, we introduce a clustering-based sampling that reorganizes the generalized dataset into balanced clusters, reducing gradient bias and improving representation diversity. For personalization, we mix a small subset of generalized samples with user-specific data to prevent catastrophic forgetting while maintaining adaptability.

While distillation and fine-tuning are well-known techniques, our work uniquely integrates them into a single solution to train a tiny model with only 281K parameters for accurate personalized gaze estimation. During distillation, we compress a large model trained on ImageNet-1K while preserving its generalizability. During fine-tuning, we personalize the model using both clustered general training samples and personalized data to avoid catastrophic forgetting. To the best of our knowledge, this is the first work to apply these techniques for training tiny personalized gaze estimation models.

2. RELATED WORK

Knowledge Distillation. Knowledge distillation compresses models by transferring knowledge from a deep teacher network to a lightweight student, enhancing inference speed without compromising performance. It is categorized into logit distillation [9, 10] and intermediate representation distillation [11–13], with our approach focusing on the latter to minimize feature discrepancies while reducing model size. FitNets [11] introduced this concept, CRD [12] applies contrastive learning for structured transfer, and DMAE [13] refines it by aligning features across different architectures using masked inputs. We construct a compact student by directly downsizing the teacher network, transferring fundamental weights to preserve key insights, and employing decoders to efficiently reconstruct features.

Personalized Gaze Estimation. Personalized gaze estimation [6, 14–16] adapts predictions to individual variations using a minimal set of personal images for precise gaze mapping. Existing personalized gaze estimation methods, such as SAGE [14], employ unsupervised adaptation, while TPGaze [6] uses meta-learning for efficient fine-tuning. Our approach leverages a distilled model, enabling efficient adaptation by fine-tuning only a small set of Adapter parameters, making it well-suited for online fine-tuning on edge devices. Additionally, clustered gaze images help mitigate catastrophic forgetting and enhance adaptation, even with as few as five user-specific images.

3. OUR APPROACH

We aim to develop a compact gaze estimation model that is both efficient and accurate for generalized and personalized tasks. Existing models are often large and prone to overfitting, especially when fine-tuning on limited personal data. To address this, we use knowledge distillation and clustering-based adaptation. First, we distill a student model from a larger teacher model, preserving essential features while reducing the model size by 10x. The student is trained on ImageNet-1K to capture broad visual representations (See Section 3.1). Given a dataset $\mathcal{D} = \{(\mathcal{X}) \mid \mathcal{X} \in \mathcal{I}_{\mathcal{D}}\}$, the student reconstructs the original images $\mathcal{X} \in \mathbb{R}^{H \times W \times 3}$ and aligns its intermediate features $\mathbf{f}^{S}$ with the teacher's $\mathbf{f}^{T} \in \mathbb{R}^{H_l \times W_l \times C_l}$. The reconstructed image is $\hat{\mathcal{X}} \in \mathbb{R}^{H \times W \times 3}$, where H, W are image dimensions and H_l, W_l, C_l represent feature map dimensions at l-th stage. We then enhance the distilled model with Adapters to form DFT Gaze for gaze prediction. To improve generalization, we apply clustering to structure the generalized dataset as $\mathcal{G} = \bigcup_g \mathcal{G}_g$, where each cluster is defined as $\mathcal{G}_g = \{(x_i^g, y_i^g) \mid x_i^g \in \mathcal{I}_{\mathcal{G}}, y_i^g \in \mathcal{Y}_{\mathcal{G}}\}_{i=1}^{N_g}\}$. Here, x_i^g and y_i^g are image-label pairs belonging to the clustered generalized set. For personalized gaze estimation, each user's dataset D_P is defined as $D_P = \{(x_i^p, y_i^p) \mid x_i^p \in \mathcal{I}_P, y_i^p \in \mathcal{Y}_P\}_{i=1}^{N_{P_j}}$, where x_i^p, y_i^p are image-label pairs for the j-th user. We use $N_{P_j} = 5$ for fine-tuning and include n_G (≈100) additional images from $\mathcal{G}$ to prevent catastrophic forgetting.

3.1. Pretraining a compact model for gaze estimation

Gaze estimation for real-time applications requires both accuracy and efficiency, but many existing models [6–8, 17] are too computationally expensive. Prior works have explored ResNet-based [6, 7] and Transformer-based [17] architectures, which achieve better accuracy but are impractical for edge devices. ConvNeXt V2-A [18] offers strong visual representations with a streamlined design, making it a more promising candidate. However, it is still too large for efficient deployment. Thus, we reduce its size to 281K parameters while preserving essential feature through knowledge distillation. Using a masked autoencoder (MAE) [18, 19] with knowledge distillation (KD), we transfer essential features from ConvNeXt V2-A into a small student to achieve high compression without compromising performance.

Unlike conventional compression methods that rely on aggressive pruning or low-rank approximations, our approach simplifies compression by directly reducing channel dimensions in later stages while keeping the first-stage channels unchanged. Each stage consists of several ConvNeXt V2 Blocks that process different resolution levels and feature abstractions, early stages capture fine details, while later stages extract high-level representations. To achieve high compression, we include a MAE during KD, where the student reconstructs missing information from the teacher's feature maps and input images. This dual reconstruction helps the student model extract essential features from limited information, improving generalization despite its compact size. By predicting both image structures and high-level representations, we maximize knowledge retention while minimizing parameter count.

During knowledge distillation, the student model takes masked input images from ImageNet-1K [20] and reconstructs both the original images $\mathcal{X}$ and teacher's intermediate features $\mathbf{f}^T$. The reconstructed image is denoted as $\hat{\mathcal{X}}$. The teacher model processes the same images without masking and provides complete feature representations. The student learns to predict missing information in its feature maps and align its outputs with the teacher's intact features (Figure 2).

We choose ImageNet-1K instead of a gaze dataset for pretraining because gaze datasets have several limitations. They are often collected in controlled environments with uniform lighting, simple backgrounds, and fixed poses, which makes it difficult for models to handle real-world variations. They also lack diverse noise, occlusions, and complex environmental factors, which reduces robustness. In addition, gaze datasets focus mainly on eye regions and facial structures, limiting their ability to learn broader visual features like edges, textures, and object shapes. Masked autoencoders need diverse training samples to reconstruct missing parts, but the low diversity in gaze datasets leads to poor or biased reconstructions. Pretraining on ImageNet-1K, which includes a wide range of textures, lighting conditions, and objects, allows our model to learn strong general features that support effective fine-tuning for gaze estimation tasks.

We reconstruct high-level features in the last two stages (l-th stage, where $l \in 3, 4$) of ConvNeXt V2-A while keeping the teacher's weights in the first stage. This approach helps the student model build on fundamental knowledge so it can develop and process abstract concepts similarly to the teacher. Each reconstruction task uses a separate ConvNeXt V2 Block [18] as a decoder, one for input image reconstruction and another for feature reconstruction. The decoder $\Psi(\mathbf{z})$ reconstructs the teacher's intermediate features from input $\mathbf{z}$:

$$\Psi(\mathbf{z}) = \mathrm{FC}\Big(\mathbf{z} + \mathrm{Conv}_{1\times1}\big(\mathrm{GRN}(\mathrm{GELU}(\hat{\mathbf{z}}))\big)\Big), \qquad (1)$$

where $\hat{\mathbf{z}} = \mathrm{Conv}_{1\times1}(\mathrm{LN}(\mathrm{DConv}_{7\times7}(\mathbf{z})))$. GRN stands for Global Response Normalization, GELU is an activation function, LN refers to Layer Normalization, and DConv represents Depthwise Convolution. We align the student's features, $\mathbf{f}_l^S$, with the teacher's features, $\mathbf{f}_l^T$, at the same stage using this decoder. The reconstruction loss for both input image and intermediate feature alignment is expressed as:

$$\begin{aligned}\mathcal{L}_{recon} = &\frac{1}{\phi(\mathcal{X}_K)} \sum_{k\in K} (\mathcal{X}_k - \hat{\mathcal{X}}_k)^2 + \\ &\gamma \sum_{l\in\{3,4\}} \frac{1}{\phi(\mathbf{f}_{l,K}^T)} \sum_{k\in K} \left(\mathbf{f}_{l,k}^T - \Psi(\mathbf{f}_{l,k}^S)\right)^2,\end{aligned} \qquad (2)$$

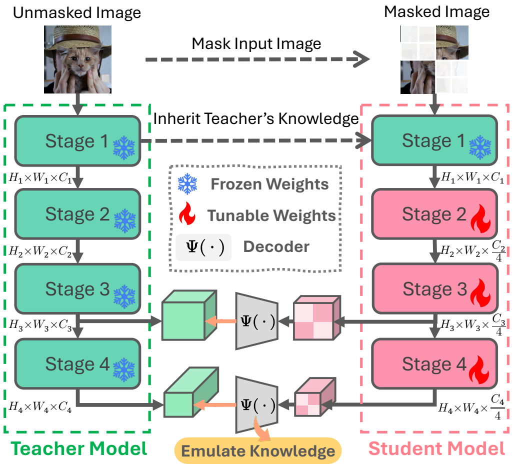
Unmasked Image
Mask Input Image
Masked Image
Inherit Teacher's Knowledge
Stage 1
Stage 2
Stage 3
Stage 4
$H_1\times W_1\times C_1$
$H_2\times W_2\times C_2$
$H_3\times W_3\times C_3$
$H_4\times W_4\times C_4$
$H_2\times W_2\times \frac{C_2}{4}$
$H_3\times W_3\times \frac{C_3}{4}$
$H_4\times W_4\times \frac{C_4}{4}$
Frozen Weights
Tunable Weights
$\Psi(\cdot)$ Decoder
$\Psi(\cdot)$
$\Psi(\cdot)$
Teacher Model
Emulate Knowledge
Student Model

Fig. 2. We downsize ConvNeXt V2-A into a student network by inheriting first-stage weights and reducing channel dimensions by 4× in stages 2-4. Each stage refines features from fine details to high-level abstractions. The student processes masked inputs, while the teacher uses unmasked ones. Separate decoders reconstruct input images and teacher features for knowledge transfer. While this diagram focuses on feature reconstruction, image reconstruction is also involved.

where K denotes the masked pixels, $\phi(\cdot)$ represents their total count, and $\gamma = 0.5$ balances the losses between image and feature reconstruction.

3.2. Clustering for adaptive gaze estimation

Building on the distilled model with rich visual priors and a wide range of visual structures, we integrate Adapters to form the DFT Gaze model, which learns gaze patterns using gaze datasets. Existing gaze estimation models struggle to balance generalization across diverse users while maintaining adaptability to individual gaze patterns. Generalized models tend to overfit to dominant gaze directions, making them less responsive to rare gaze variations. Conversely, fine-tuning on small personalized datasets often results in overfitting, erasing previously learned knowledge and reducing the model's ability to generalize beyond a user's typical gaze patterns. This tradeoff limits the effectiveness of gaze estimation models across diverse users and adaptability for personalization.

To overcome these challenges, we introduce lightweight Adapters coupled with a clustering-based training mechanism. First, Adapters are integrated into the distilled model and trained on a clustered generalized gaze dataset ($\mathcal{G}$) to improve diversity in gaze representation. Clustering prevents frequent gaze directions from dominating the learning process by ensuring balanced sampling, reducing gradient bias, and improving feature diversity. Once the model learns generalized gaze estimation, the same Adapters are refined for personalization.

However, catastrophic forgetting can occur when fine-tuning on small datasets (e.g., 5 personal images) [6, 15, 21, 22]. If the model overfits and overwrites previously learned general knowledge, it may generalize poorly. To mitigate this, instead of fine-tuning solely on limited user-specific data, we also incorporate a small subset of the clustered generalized gaze dataset. This preserves a broad feature space, stabilizes adaptation, and ensures the model retains general knowledge while adapting to individual gaze behaviors.

In generalized gaze estimation, the DFT Gaze model learns diverse gaze variations using Adapters, which consist of two fully connected (FC) layers, BatchNorm (BN), and LeakyReLU (LReLU) activation. Only the Adapters are fine-tuned, while the rest of the model remains unchanged to preserve learned visual knowledge. However, a major challenge in generalized gaze estimation is the dominance of common gaze directions. Frequent gaze patterns, such as looking straight ahead, appear more often in the dataset, leading the model to focus disproportionately on them while struggling to recognize less frequent gaze variations, such as extreme side gazes. To address this, we apply K-means clustering to divide the generalized dataset (D_G) into 15 balanced groups, forming the clustered generalized set ($\mathcal{G}$). Even sampling from each group ensures better representation of rare gaze directions and reduces gradient bias during training. Adapters within each ConvNeXt V2 Block adjust internal features to better align with varied gaze patterns, enhancing the model's robustness. The transformation within an Adapter is defined as:

$$\text{Adapter}(\mathbf{f}^V) = \text{FC}_{\text{up}}\Big(\text{LReLU}\big(\text{BN}(\text{FC}_{down}(\mathbf{f}^V))\big)\Big) \quad (3)$$

Here, $\mathbf{f}^V$ denotes the features from the final convolutional layer of each block. The FC_{down} layer compresses these features to a quarter of their original dimension, focusing on the most relevant attributes. The FC_{up} layer then restores the features back to their original dimensions.

In personalized gaze estimation, the model is fine-tuned to adapt to individual users. This involves refining Adapters in the last two stages of the DFT Gaze model using a personalized dataset (D_P) with the first five gaze images per participant. A key challenge is limited diversity in user data. Since gaze patterns are often repetitive, fine-tuning on a small dataset risks overfitting, making the model too specialized and less effective for unseen gaze directions. For example, if a user frequently looks slightly left or right, the model may struggle with rare gaze shifts like extreme upward glances. To prevent this, we introduce a small subset of the clustered generalized dataset ($\mathcal{G}$) during personalization. This helps the model adapt to user-specific patterns while retaining broader gaze variations. Additionally, incorporating samples from $\mathcal{G}$ mitigates catastrophic forgetting, preserving knowledge from generalized training.

To optimize the DFT Gaze model, we use an L1 loss function for both generalized and personalized gaze estimation.

Table 1. Comparison of state-of-the-art methods for generalized and personalized gaze estimation. Bold indicates the best performance; italics denote the second-best.

Model	#Param (Tunable)	MPIIGaze	MPIIFaceGaze	AEA	OpenEDS2020
Generalized Gaze Estimation					
GazeNet [8]	90.24M (90.24M)	*5.70*	5.76	3.01	*7.51*
ConvNeXt V2-A (Teacher model) [18]	3.6M (191.7K)	**5.30**	**4.29**	**1.94**	**6.90**
DFT Gaze	**281K (14.43K)**	6.13	*5.17*	*2.14*	7.82
Personalized Gaze Estimation					
GazeNet [8]	90.24M (90.24M)	**5.39**	-	4.16	6.57
PNP-GA [7]	119.5M (116.9M)	-	6.91	-	-
RUDA [23]	12.20M (12.20M)	-	6.86	-	-
TPGaze [6]	11.82M (125K)	-	6.30	-	-
ConvNeXt V2-A (Teacher model) [18]	3.6M (191.7K)	*5.49*	**4.60**	**2.32**	**5.36**
DFT Gaze	**281K (14.43K)**	6.61	*5.35*	*2.60*	*5.80*

The generalized dataset is divided into clusters, each containing N_g samples. Given an input $x_i^g \in \mathbb{R}^{H \times W \times 3}$ from $\mathcal{G}$, the model predicts $\hat{y}_i^g \in \mathbb{R}^2$, compared against the ground truth $y_i^g \in \mathbb{R}^2$. For personalization, the dataset includes user-specific images $x_i^p \in \mathbb{R}^{H \times W \times 3}$ and a small subset of generalized samples x_i^g from $\mathcal{G}$ to prevent catastrophic forgetting. The model predicts $\hat{y}_i^p \in \mathbb{R}^2$, with labels $y_i^p \in \mathbb{R}^2$. Both tasks are optimized separately.

$$\mathcal{L}_G = \frac{1}{N_G} \sum_{i=1}^{N_G} |y_i^g - \hat{y}_i^g| \quad (4)$$

where N_G is the total number of generalized samples.

$$\mathcal{L}_{P_j} = \frac{1}{N_{P_j} + n_G} \sum_{i=1}^{N_{P_j}+n_G} |y_i^p - \hat{y}_i^p| \quad (5)$$

where N_{P_j} is the number of user-specific samples for the j-th user, and n_G is the number of additional generalized samples.

4. EXPERIMENTS

We evaluate DFT Gaze on four benchmarks and test personalization on unseen users not included in the generalized test set. Since the personalized data has a different distribution, gaze error may be higher, which aligns with real-world adaptation challenges.

4.1. Experimental setup and implementation details

Both DFT Gaze and ConvNeXt V2-A are trained using AdamW. For generalized gaze estimation, we use a batch size of 64. Models are trained for 100 epochs on MPIIGaze and MPIIFaceGaze, and for 200 epochs on AEA and OpenEDS2020. Learning rate is set to 6.25×10^{-5} for all datasets. For personalized gaze estimation, we use a batch size of 8 and train for 100 epochs on each of the datasets. Learning rate is set to 1×10^{-3} for MPIIGaze and MPIIFaceGaze, and 1×10^{-5} for AEA and OpenEDS2020.

4.2. Comparison with state-of-the-art methods

Table 1 presents the performance comparison of DFT Gaze against state-of-the-art methods and the teacher model for

Table 2. Impact of imbalanced data on generalized gaze estimation and catastrophic forgetting in personalized estimation.

Sampling Method	OpenEDS2020
Generalized Gaze Estimation	
w/ clustered generalized set ($\mathcal{G}$)	**7.82**
w/o clustered set (imbalanced gaze data)	10.23
Personalized Gaze Estimation	
w/ small subset of clustered generalized set ($\mathcal{G}$)	**5.80**
w/o clustered set (catastrophic forgetting)	6.92

Table 3. Impact of teacher knowledge, feature reconstruction.

Distillation Method	OpenEDS2020
Generalized Gaze Estimation	
Full generalized DFT Gaze	**7.82**
w/o inheriting teacher's knowledge	9.50
w/o reconstructing teacher's features	10.52

both generalized and personalized gaze estimation. DFT Gaze achieves the smallest parameter count (281K) and the lowest number of tunable parameters (14.43K) while maintaining minimal gaze error increase compared to the teacher model (ConvNeXt V2-A) in generalized gaze estimation. We note that PnP-GA, RUDA, and TPGaze address a more difficult unsupervised adaptation setting from ETH-XGaze [24], while GazeNet, ConvNeXt V2-A, and DFT Gaze are trained in-domain with supervision.

4.3. Ablation study

Cluster-based gaze adaptation. Table 2 shows generalized set clustering reduces overfitting to frequent gaze angles, improving generalization. For personalized gaze estimation, fine-tuning with a small clustered subset prevents catastrophic forgetting and preserves diversity, while direct fine-tuning results in greater error from overfitting.

Teacher knowledge and feature reconstruction. Table 3 shows the impact of teacher knowledge and feature reconstruction. DFT Gaze achieves the lowest error of $7.82°$. Omitting knowledge inheritance increases error to $9.50°$. Removing feature reconstruction further increases error to $10.52°$.

Reconstructing stages. Table 4 shows that reconstructing stages 3 and 4 achieves the lowest error ($7.82°$). Limiting reconstruction to stage 4 increases the error to $8.92°$, showing the importance of earlier stages.

Adapter designs. Table 5 shows that Adapter design impacts performance. DFT Gaze, with two fully connected layers (281K parameters, 14.43K tunable), achieves the lowest error, while a single-layer projection increases error.

Adapter channel reduction. Table 6 shows the trade-off between model size and performance with channel reduction in Adapters. DFT Gaze uses a 4× reduction, maintaining efficiency while preserving essential features. Further reducing to 8× compromises performance.

Table 4. Impact of reconstruction stages.

Reconstructed Stages	OpenEDS2020
Generalized Gaze Estimation	
Stages 3 and 4	**7.82**
Stage 4 only	8.92

Table 5. Impact of Adapter design.

Adapter Design	#Params (Tunable)	OpenEDS2020
Generalized Gaze Estimation		
Full generalized DFT Gaze	281K (14.43K)	**7.82**
Single FC layer	293.82K (27.24K)	10.86

Table 6. Impact of Adapter channel reduction.

Adapter Channel Reduction	#Params (Tunable)	OpenEDS2020
Generalized Gaze Estimation		
No reduction	321.8K (55.2K)	8.43
2× reduction	294.6K (28.0K)	8.75
4× reduction	281K (14.43K)	**7.82**
8× reduction	274.2K (7.6K)	10.92

4.4. Gaze estimation latency on edge device

To evaluate real-time gaze estimation, we measured the latency of ConvNeXt V2-A (teacher), DFT Gaze (student), GazeNet, and TPGaze on a Raspberry Pi 4 (8GB RAM) over 1,000 iterations using the AEA dataset. As shown in Figure 3, GazeNet had the highest latency (1960.80 ms), followed by ConvNeXt V2-A (744.68 ms) and TPGaze (560.25 ms). DFT Gaze achieved the lowest latency (360.21 ms), making it the best choice for real-time edge applications.

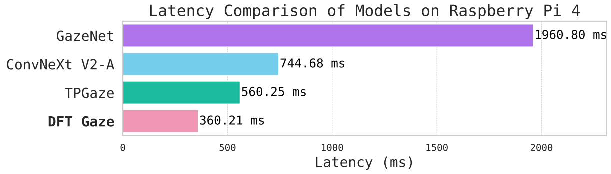

Fig. 3. Latency comparison of gaze models on the AEA dataset: DFT Gaze is the fastest (≈360 ms), followed by TPGaze (≈560 ms), ConvNeXt V2-A (≈745 ms), and GazeNet (≈1961 ms).

5. CONCLUSION

We introduce DFT Gaze, an ultra-compact and efficient model for personalized gaze estimation. Through structured knowledge distillation with masked autoencoders, our approach learns rich visual representations from a large teacher model. Moreover, clustering-based training preserves diverse gaze patterns and prevents catastrophic forgetting. We improve adaptability across different users. Despite having just 281K parameters, DFT Gaze outperforms state-of-the-art methods in accuracy and efficiency, making it well-suited for real-time personalized gaze estimation on edge devices.

References

[1] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belén Masiá, and Gordon Wetzstein, “Saliency in VR: how do people explore virtual environments?,” *TVCG*, vol. 24, no. 4, pp. 1633–1642, 2018.

[2] Alisa Burova, John Mäkelä, Jaakko Hakulinen, Tuuli Keskinen, Hanna Heinonen, Sanni Siltanen, and Markku Turunen, “Utilizing VR and gaze tracking to develop AR solutions for industrial maintenance,” in *CHI*, 2020, pp. 1–13.

[3] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu, “Conversational gaze aversion for human-like robots,” in *HRI*, 2014, pp. 25–32.

[4] Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong, “Stressclick: Sensing stress from gaze-click patterns,” in *MM*. 2016, pp. 1395–1404, ACM.

[5] Carlos Hitoshi Morimoto and Marcio R. M. Mimica, “Eye gaze tracking techniques for interactive applications,” *Computer vision and image understanding*, vol. 98, no. 1, pp. 4–24, 2005.

[6] Huan Liu, Julia Qi, Zhenhao Li, Mohammad Hassanpour, Yang Wang, Konstantinos N. Plataniotis, and Yuanhao Yu, “Test-time personalization with meta prompt for gaze estimation,” in *AAAI*, 2024, pp. 3621–3629.

[7] Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu, “Generalizing gaze estimation with outlier-guided collaborative adaptation,” in *ICCV*, 2021, pp. 3815–3824.

[8] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling, “Mpiigaze: Real-world dataset and deep appearance-based gaze estimation,” *TPAMI*, vol. 41, no. 1, pp. 162–175, 2019.

[9] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh, “Improved knowledge distillation via teacher assistant,” in *AAAI*, 2020, pp. 5191–5198.

[10] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang, “Decoupled knowledge distillation,” in *CVPR*, 2022, pp. 11943–11952.

[11] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio, “Fitnets: Hints for thin deep nets,” in *ICLR*, 2015.

[12] Yonglong Tian, Dilip Krishnan, and Phillip Isola, “Contrastive representation distillation,” in *ICLR*, 2020.

[13] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan L. Yuille, Yuyin Zhou, and Cihang Xie, “Masked autoencoders enable efficient knowledge distillers,” in *CVPR*, 2023, pp. 24256–24265.

[14] 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 *ICCVW*, 2019, pp. 1149–1158.

[15] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz, “Few-shot adaptive gaze estimation,” in *ICCV*, 2019, pp. 9367–9376.

[16] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez, “A differential approach for gaze estimation,” *IEEE TPAMI*, vol. 43, no. 3, pp. 1092–1099, 2021.

[17] Yihua Cheng and Feng Lu, “Gaze estimation using transformer,” in *ICPR*, 2022, pp. 3341–3347.

[18] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie, “Convnext V2: co-designing and scaling convnets with masked autoencoders,” in *CVPR*, 2023, pp. 16133–16142.

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick, “Masked autoencoders are scalable vision learners,” in *CVPR*, 2022, pp. 15979–15988.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*, 2009, pp. 248–255.

[21] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory G. Slabaugh, and Tinne Tuytelaars, “Continual learning: A comparative study on how to defy forgetting in classification tasks,” 2019.

[22] Johannes Schneider and Michail Vlachos, “Personalization of deep learning,” *Proceedings of the 3rd International Data Science Conference–iDSC2020*, pp. 89–96, 2021.

[23] Yiwei Bao, Yunfei Liu, Haofei Wang, and Feng Lu, “Generalizing gaze estimation with rotation consistency,” in *CVPR*, 2022, pp. 4197–4206.

[24] 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, vol. 12350, pp. 365–381.