Rapidly Deploying On-Device Eye Tracking by Distilling Visual Foundation Models

Cheng Jiang[1,2⋆], Jogendra Kundu[1], David Colmenares[1], Fengting Yang[1], Joseph P Robinson[1], Ali Behrooz[1], and Yatong An[1⋆⋆]

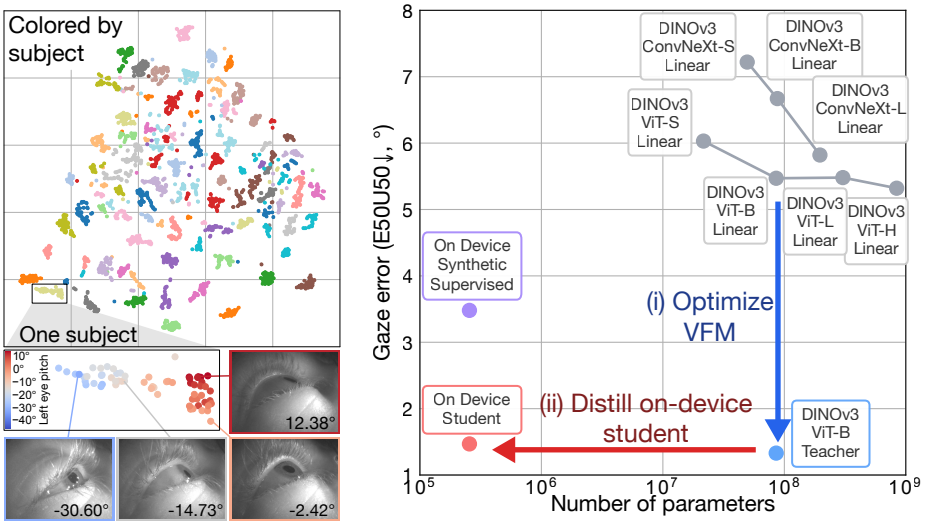
[1] Meta Reality Labs, USA
[2] University of Michigan, USA

Abstract. Eye tracking (ET) plays a critical role in augmented and virtual reality applications. However, rapidly deploying *high-accuracy*, on-device gaze estimation for new products remains challenging because hardware configurations (e.g., camera placement, camera pose, and illumination) often change across device generations. Visual foundation models (VFMs) excel on natural-image benchmarks and offer a promising path to rapid training and deployment; yet, we find that off-the-shelf VFMs still struggle to reach high accuracy on specialized near-eye infrared images. To close this gap, we introduce *DistillGaze*, a framework that distills a VFM using labeled synthetic data and unlabeled real data for rapid, high-accuracy on-device gaze estimation. DistillGaze proceeds in two stages. First, we adapt a VFM into a domain-specialized teacher using synthetic gaze labels and unlabeled real images. Synthetic data provide scalable, high-quality gaze supervision, while unlabeled real data bridges the synthetic-to-real domain gap. Second, we train an on-device student from both teacher guidance and self-training. Evaluated on a large-scale crowd-sourced dataset spanning more than 2,000 participants, DistillGaze reduces median gaze error by 58.6% relative to synthetic-only baselines while maintaining a lightweight 256K-parameter model suitable for real-time on-device deployment. More broadly, DistillGaze offers an efficient path to training and deploying ET models that adapt to hardware changes, and a recipe for combining synthetic supervision with unlabeled real data in on-device regression tasks.

1 Introduction

Eye tracking (ET) is a core sensing capability for augmented and virtual reality (AR/VR) applications, enabling foveated rendering, gaze-driven interaction, and attention-aware interfaces [26]. However, building a production-grade ET system is costly and time-consuming. Conventionally, teams must iterate through hardware prototyping, calibration, data collection, and annotation to obtain ground-truth gaze supervision [29]. Each new device generation can trigger a full repetition of data acquisition and model retraining [26], due to changes

⋆ Work done during an internship at Meta.
⋆⋆ Correspondence: yatong@meta.com.



(a) DINOv3 embedding t-SNE. **(b)** Workflow overview.

Fig. 1: (a) VFM embeddings cluster by subject, obscuring gaze structure. t-SNE of DINOv3 embeddings for near-eye images from 64 randomly sampled subjects: embeddings cluster by subject rather than gaze, yet smooth within-cluster gradients correlate with gaze direction, indicating that gaze cues are present but dominated by identity. **(b) Workflow overview.** Linear probing of DINOv3 with synthetic supervision (gray) substantially underperforms our on-device baseline (purple), despite having orders of magnitude more parameters, motivating a two-stage approach: (i) optimize a ViT-B teacher from DINOv3; (ii) distill to an on-device student.

in hardware configuration such as differences in sensor characteristics, camera placement, optics, illumination or on-device image processing. Moreover, gaze supervision itself is often unreliable: even controlled protocols suffer from imperfect fixation and incomplete compliance, producing label noise that is difficult to detect at scale [29].

As AR/VR hardware iterations accelerate, ET models must be deployed on each new device without waiting for a full labeled-data collection cycle, which for crowd-sourced protocols takes *weeks to months* per device variant. Synthetic data offers a practical path: from a device's camera intrinsics/extrinsics and illumination geometry, photo-realistic near-eye images with pixel-perfect gaze labels can be rendered at scale across diverse eye appearances and gaze directions. Because existing 3D eye assets and subject appearance models are reusable across generations, only the known optical parameters need updating, cutting turnaround from *months to days*. Rendered images, however, do not perfectly match the real target distribution, leaving residual gaps in texture, noise characteristics, and optical artifacts. Unlabeled real images help close this gap at minimal cost: they capture target-device appearance statistics and can be collected passively during normal use, before a calibrated ET system exists. We therefore pair la-

beled synthetic data with unlabeled real images for *rapid, label-free adaptation* to each new device.

Visual foundation models (VFMs) such as DINOv3 [38] and SAM3 [6] have substantially improved representation quality and transfer on natural-image benchmarks, suggesting a path to bypass costly labeled-data collection for eye tracking. These gains, however, are less consistent for specialized imaging modalities, where both the sensing modality and downstream objective differ from web images. This has prompted domain-specific foundation models for medical imaging [8, 23], remote sensing [19, 27], and autonomous driving [41]. On-device ET poses the same challenge: images are captured in near-infrared at close range and contain sensor-, device-, and subject-specific nuances, and high accuracy demands precise regression of subtle appearance variations in the pupil, iris, and eyelids rather than high-level semantic understanding. Fig. 1 illustrates this limitation. Linear probing of off-the-shelf DINOv3 features (Fig. 1b, gray) yields gaze errors exceeding 5°, far short of task-specific on-device baselines trained on labeled synthetic data (purple) and of on-device ET requirements. The t-SNE visualization in Fig. 1a explains the gap: representations cluster tightly by subject, with gaze subordinated to identity. Yet, the smooth within-cluster gradients that correlate with gaze direction show that VFMs do encode a usable gaze signal; extracting it requires in-domain adaptation to reshape representations toward gaze-relevant features.

These observations motivate *DistillGaze*, a framework that optimizes a VFM for rapid, label-free ET deployment. Our approach proceeds in two stages (Fig. 1b). In the first stage (blue arrow), we adapt an off-the-shelf VFM to the eye tracking domain using labeled synthetic images and unlabeled real data, without any real ground-truth gaze labels, to bridge the gap between web-scale pretraining and task-specific ET requirements. In the second stage (red arrow), we distill the optimized model into a lightweight student with two complementary objectives: representation alignment with the fine-tuned VFM teacher and consistency regularization via an exponential moving average (EMA) of the student. The distilled model preserves high-quality ET performance while remaining suitable for real-time on-device deployment.

We evaluate DistillGaze on a large-scale ET dataset collected with Project Aria glasses [29] from over 2,000 crowd-sourced participants under in-the-wild conditions, using no ground-truth gaze labels on real data. Our optimized VFM improves both median and tail error over synthetic-only baselines: E50U50 (*i.e.*, the median gaze error of the median subject) drops from 3.48° to 1.33° (↓61.8%), and E90U90 (*i.e.*, the 90th-percentile error of the 90th-percentile users) drops from 14.84° to 8.32° (↓43.9%). The final distilled model achieves 1.44° E50U50 and 8.45° E90U90, 58.6% and 43.1% better than the synthetic-supervised baseline, with just 256K parameters, suitable for real-time on-device deployment.

The main contributions of this paper are as follows:

- We analyze off-the-shelf DINOv3 representations via t-SNE and find that embeddings cluster by subject, which limits direct transfer performance

on ET. However, intra-subject distributions exhibit smooth gaze-correlated structure, indicating strong potential for adaptation.
- We propose *DistillGaze*, a two-stage framework that adapts a VFM using synthetic labels and unlabeled real data, without any real gaze labels, then distills it into a compact student for on-device deployment.
- We achieve state-of-the-art results on a large-scale Project Aria eye tracking benchmark, with significant gains over synthetic-supervised baselines with no increase in model size.

2 Related Work

2.1 ET in AR/VR

ET is essential for AR/VR applications. Typically, ET takes a near-eye image as input and predicts the direction of gaze (*i.e.* pitch and yaw angles). Eye tracking methods generally fall into two categories: geometry-based and learning-based. Geometric methods assume an explicit eye model (e.g., sphere-on-sphere) [14, 28, 51] and optimize gaze parameters to align model predictions with image observations. While interpretable and efficient, these methods often make simplifying assumptions about eye shape, struggling with large gaze angles and occlusion.

Learning-based methods are now dominant, with CNN and transformer architectures achieving improved robustness and accuracy. However, these methods are typically fully supervised, requiring large amounts of annotated gaze data to generalize across users and conditions. Much prior work has been developed for webcams or screen-mounted cameras that capture near-frontal eye views under relatively controlled illumination [10,11,20,48,49], where large-scale data collection is comparatively straightforward. In contrast, head-mounted infrared (IR) cameras in AR/VR headsets operate at extreme off-axis angles with a limited field of view, producing images with perspective distortion, partial iris visibility, and strong specularities due to active IR illumination [12, 22, 29, 33, 45], making both data collection and generalization substantially more challenging. Compounding this difficulty, even controlled calibration protocols suffer from imperfect fixation and incomplete user compliance, introducing supervision noise that is difficult to detect at scale [29].

To address data acquisition and label quality challenges, synthetic data and domain adaptation methods have gained interest in recent years. Synthetic eye images [34, 42] can be generated at scale, covering subject diversity and imaging conditions that are impractical to capture with calibrated ground truth [26]. Yet, the synthetic-to-real domain gap limits transfer performance, which caps accuracy when training relies solely on rendered data. Domain adaptation offers a complementary strategy to bridge the synthetic-to-real gap [22,25,31,37,42,43]. More recently, self-supervised learning has emerged as a paradigm for training encoders on unlabeled data, improving cross-domain generalization [1, 46].

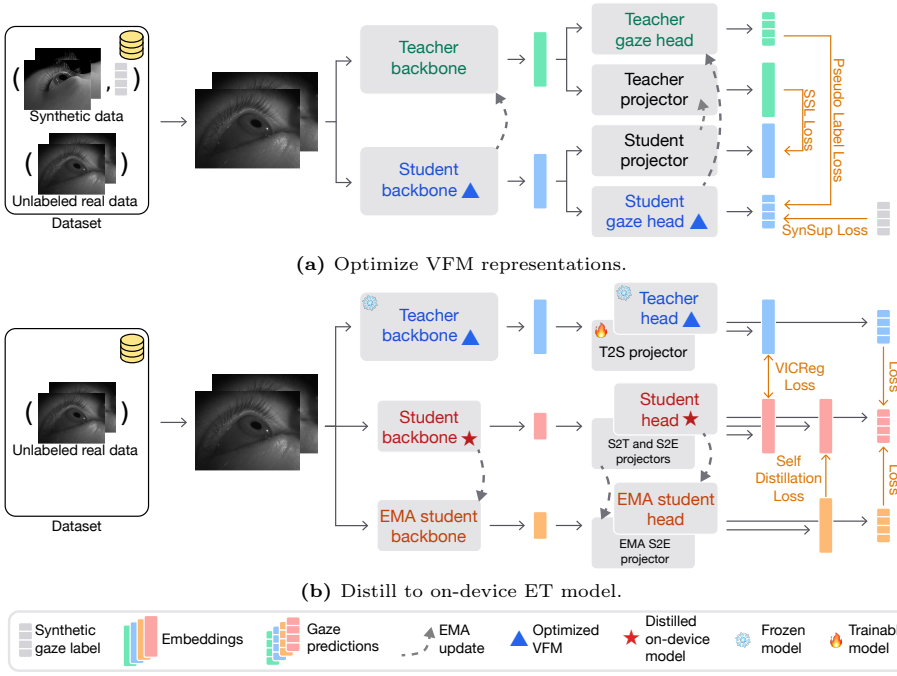
2.2 Self-Supervised Representation Learning

Self-supervised learning (SSL) learns representations by encouraging invariance across augmented views of unlabeled data. Modern SSL methods can be categorized into four broad families [3]: contrastive learning, self-distillation, canonical correlation analysis, and masked image modeling. Early contrastive methods [9, 17] require large batch sizes or memory banks, limiting scalability. Self-distillation methods typically adopt a student-teacher framework in which the teacher is an exponential moving average (EMA) of the student to prevent representational collapse [7, 13, 32, 38], and have achieved state-of-the-art performance on a range of benchmarks. Canonical-correlation-analysis-based methods, such as [5], avoid representational collapse through variance-invariance-covariance regularization without imposing a categorical structure on the embedding space. Masked image modeling methods [4, 16, 32, 50] take inspiration from language modeling and predict masked tokens from images. Most SSL methods are evaluated on classification or segmentation benchmarks, and many use loss functions that encourage a soft clustering of representations. When extended to large web-scale datasets, these techniques yield visual foundation models (VFMs). Although SSL has transformed representation learning for natural images, its application to near-eye images and gaze estimation remains underexplored.

2.3 Visual Foundation Models and Knowledge Distillation

Foundation models pretrained on large-scale data exhibit strong transfer to diverse downstream tasks. CLIP [35] aligns visual and language representations through contrastive pretraining on web-scale image-text pairs. SAM3 [6] provides segmentation conditioned on text and visual prompts. Sapiens2 [21] introduces a family of human-centric foundation models, trained on 1-billion images, to learn human appearance and semantics by combining masked reconstruction and contrastive objectives. DINOv3 [38] scales self-supervised learning to a family of models of different sizes, producing features with improved semantic structure. Despite success on natural images, we find these models transfer poorly to near-eye gaze estimation due to the significant domain shift between web-scale RGB images and the infrared, off-axis imagery of AR/VR headsets.

Beyond domain shift, standard VFMs require substantial computational resources to achieve their best performance, rendering them impractical for on-device deployment even at the smallest model size within their family. Knowledge distillation (KD) offers a natural remedy by transferring learned representations from a larger teacher to a compact student network. Early KD methods [18, 36, 39, 40, 47] laid the foundation for modern approaches by using soft target logits, intermediate feature matching, attention map alignment, relational structure preservation, and contrastive objectives. Community KD [24] trains multiple online students jointly through co-distillation with a pretrained teacher, allowing for bidirectional knowledge transfer. We build on Community KD, adapting it to combine supervision from an optimized teacher and self-distillation of the student.



(a) Optimize VFM representations.

(b) Distill to on-device ET model.

Fig. 2: Overview of DistillGaze. (a) We optimize a VFM using synthetic supervision and self-distillation on unlabeled real data. **(b)** The optimized VFM is distilled to a lightweight student with 256K parameters, with an EMA student providing complementary supervision. Only the student is deployed.

3 Methods

Fig. 2 illustrates our two-stage approach. We first describe our base ET architecture (Sec. 3.1), then detail our VFM optimization strategy (Sec. 3.2), and finally present the distillation procedure (Sec. 3.3).

3.1 Preliminaries: ET Gaze Estimation Model

Our gaze estimation architecture follows [29], mapping a synchronized pair of eye images to a per-eye gaze direction. Given input images $\boldsymbol{X} = (\boldsymbol{X}^L, \boldsymbol{X}^R)$ for the left and right eyes, a shared-weight backbone f extracts per-eye features:

$$\boldsymbol{h}^L = f(\boldsymbol{X}^L), \quad \boldsymbol{h}^R = f(\boldsymbol{X}^R), \tag{1}$$

which are concatenated to form a binocular representation $\boldsymbol{h} = [\boldsymbol{h}^L; \boldsymbol{h}^R]$. A fully connected regression head g predicts per-eye gaze rays:

$$\hat{\boldsymbol{y}} = g(\boldsymbol{h}) = \left[\hat{\theta}^L, \hat{\varphi}^L, \hat{\theta}^R, \hat{\varphi}^R\right] \in \mathbb{R}^4, \tag{2}$$

where θ and φ denote yaw and pitch angles, respectively.

Following [29], we supervise gaze predictions using a smooth L1 loss with outlier rejection, which provides robustness to annotation noise:

$$\mathcal{L}_{\text{Gaze}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \begin{cases} \frac{1}{2}\frac{(\boldsymbol{y}-\hat{\boldsymbol{y}})^2}{\beta} & \text{if } |\boldsymbol{y}-\hat{\boldsymbol{y}}| < \beta, \\ |\boldsymbol{y}-\hat{\boldsymbol{y}}| - \frac{1}{2}\beta & \text{if } \beta \le |\boldsymbol{y}-\hat{\boldsymbol{y}}| < \gamma, \\ k(|\boldsymbol{y}-\hat{\boldsymbol{y}}| - \frac{1}{2}\beta) & \text{if } |\boldsymbol{y}-\hat{\boldsymbol{y}}| \ge \gamma, \end{cases} \tag{3}$$

where k, β, and γ are hyperparameters, and $\boldsymbol{y}$ is normalized gaze supervision. The loss is quadratic for small errors, linear for moderate errors, and scaled by $k < 1$ for large outliers beyond γ.

3.2 Optimizing VFM Representations

Foundation models pretrained on web-scale data generalize well across many vision tasks. However, frozen foundation-model features underperform a small model trained on simulated data for near-eye gaze estimation (Sec. 5.1), suggesting that their high-level representations, optimized for semantic understanding, do not transfer directly to our domain, even if low-level features remain useful.

To adapt these representations for gaze estimation, we adopt a teacher-student self-distillation framework combining synthetic supervision with pseudo-label training on unlabeled real data (Fig. 2a). Let f_t, f_s denote the initialized DINOv3 teacher and student backbones, and p_t, p_s their projectors, respectively. The teacher receives weakly augmented images $\boldsymbol{X}_w$ and the student receives strongly augmented images $\boldsymbol{X}_s$. We compute projected embeddings:

$$\boldsymbol{z}_t = p_t(\boldsymbol{h}_t) = p_t(f_t(\boldsymbol{X}_w)), \qquad \boldsymbol{z}_s = p_s(\boldsymbol{h}_s) = p_s(f_s(\boldsymbol{X}_s)). \tag{4}$$

The self-distillation loss encourages alignment between teacher and student embeddings:

$$\mathcal{L}_{\text{SD}}^{\text{VFM}} = \|\boldsymbol{z}_t - \boldsymbol{z}_s\|_2^2 \tag{5}$$

We then predict gaze using the prediction heads g_t, g_s for teacher and student models, respectively. Let $\hat{\boldsymbol{y}}_t = g_t(\boldsymbol{h}_t)$ and $\hat{\boldsymbol{y}}_s = g_s(\boldsymbol{h}_s)$ denote the teacher and student predictions, respectively. For synthetic data with labels $\boldsymbol{y}$, we apply synthetic supervised (SynSup) gaze loss:

$$\mathcal{L}_{\text{SynSup}} = \mathcal{L}_{\text{Gaze}}(\boldsymbol{y}, \hat{\boldsymbol{y}}_s). \tag{6}$$

For unlabeled real data, we use teacher predictions as pseudo-labels:

$$\mathcal{L}_{\text{Pseudo}} = \mathcal{L}_{\text{Gaze}}(\hat{\boldsymbol{y}}_t, \hat{\boldsymbol{y}}_s). \tag{7}$$

The total objective for optimizing the VFM is:

$$\begin{aligned} \mathcal{L}_{\text{OptVFM}} = {} & \lambda_{\text{SynSup}} \sum_{\boldsymbol{X} \in \mathcal{D}_{\text{syn}}} \mathcal{L}_{\text{SynSup}} \\ & + (1 - \lambda_{\text{SynSup}}) \left[\sum_{\boldsymbol{X} \in \mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{syn}}} \mathcal{L}_{\text{SD}}^{\text{VFM}} + \sum_{\boldsymbol{X} \in \mathcal{D}_{\text{real}}} \mathcal{L}_{\text{Pseudo}} \right], \end{aligned} \tag{8}$$

where $\mathcal{D}_{\text{syn}}$, $\mathcal{D}_{\text{real}}$ denote the labeled synthetic and unlabeled real datasets, respectively. The hyperparameter λ_{SynSup} balances the two sources of supervision and is annealed using a cosine schedule: training initially assigns a high weight to synthetic supervision to align the VFM representations with the gaze task, then gradually shifts to self-distillation using all available data, along with pseudo-label supervision using unlabeled real data.

For stable self-distillation, the teacher parameters are updated via EMA [7, 13]:

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s, \tag{9}$$

where θ_t and θ_s denote the teacher and student parameters, respectively, including the backbone, projector, and gaze head.

3.3 Distilling to an On-Device ET Model

With the optimized VFM from Sec. 3.2, we now distill its representations into a lightweight student network suitable for on-device deployment. As illustrated in Fig. 2b, our framework comprises three models:

- A large **teacher** f_t: our fine-tuned VFM from Sec. 3.2;
- A **student** f_s: an on-device model pretrained on synthetic data;
- An **EMA student** f_e: an exponential moving average of f_s.

Here, we reuse the teacher/student notation, but importantly, the roles of these models differ from Sec. 3.2: in the distillation stage, the teacher is the optimized VFM and the student is an on-device model. We additionally introduce an EMA student to provide complementary targets for self-distillation and pseudo-label supervision. Our design is inspired by community KD [24] with key differences: (1) we distill in both feature space (via a VIC-KD loss [15]) and prediction space (via pseudo-labels); and (2) the EMA student is updated via momentum rather than gradient descent, drawing inspiration from momentum-based self-supervised learning [7,13]. All three models share a common structure: a backbone for image feature extraction, projectors that map features into distillation spaces, and a prediction head to estimate gaze. The teacher uses the optimized VFM backbone f_t and gaze head g_t from Sec. 3.2; the student models use backbones and heads pretrained on synthetic data. The projectors are randomly initialized. During training, the teacher and the EMA student receive weakly augmented images $\boldsymbol{X}_w$, while the student receives strongly augmented images $\boldsymbol{X}_s$.

Given an image, the teacher produces embedding $\boldsymbol{h}_t = f_t(\boldsymbol{X}_w)$, projection $\boldsymbol{z}_t = p_t(\boldsymbol{h}_t)$ for distillation, and gaze prediction $\hat{\boldsymbol{y}}_t = g_t(\boldsymbol{h}_t)$. The student backbone f_s extracts features $\boldsymbol{h}_s = f_s(\boldsymbol{X}_s)$, which are mapped to projections $\boldsymbol{z}_{s\to t} = p_{s\to t}(\boldsymbol{h}_s)$ and $\boldsymbol{z}_{s\to e} = p_{s\to e}(\boldsymbol{h}_s)$ for distillation with the teacher and the EMA student, respectively. The head g_s produces gaze prediction $\hat{\boldsymbol{y}}_s = g_s(\boldsymbol{h}_s)$. The student receives gradient updates from all loss terms.

The EMA student maintains the backbone f_e, the projector p_e, and the head g_e, all updated via the exponential moving average from f_s, $p_{s\to e}$, and g_s,

respectively. This produces stable embedding $\boldsymbol{h}_e = f_e(\boldsymbol{X}_w)$, projection $\boldsymbol{z}_e = p_e(\boldsymbol{h}_e)$, and prediction $\hat{\boldsymbol{y}}_e = g_e(\boldsymbol{h}_e)$.

We train the framework with complementary objectives in both feature space and gaze prediction space. For teacher-student distillation in feature space, we use an invariance loss between teacher and student projections, along with variance and covariance regularization:

$$\mathcal{L}_{\text{KD}} = \lambda_{\text{inv}} \|\boldsymbol{z}_t - \boldsymbol{z}_{s\to t}\|_2^2 + \lambda_{\text{var}} v(\boldsymbol{z}_{s\to t}) + \lambda_{\text{cov}} c(\boldsymbol{z}_{s\to t}), \tag{10}$$

where $v(\cdot)$ and $c(\cdot)$ denote variance and covariance regularization terms [5]. λ_{inv}, λ_{var}, and λ_{cov} are hyperparameters. The teacher also provides pseudo-labels:

$$\mathcal{L}_{\text{Pseudo-}t} = \mathcal{L}_{\text{Gaze}}(\hat{\boldsymbol{y}}_t, \hat{\boldsymbol{y}}_s). \tag{11}$$

The EMA student acts as a momentum teacher for self-distillation. We use an L2 alignment loss and pseudo-label supervision:

$$\mathcal{L}_{\text{SD}}^{\text{On-device}} = \|\boldsymbol{z}_e - \boldsymbol{z}_{s\to e}\|_2^2; \qquad \mathcal{L}_{\text{Pseudo-}e} = \mathcal{L}_{\text{Gaze}}(\hat{\boldsymbol{y}}_e, \hat{\boldsymbol{y}}_s). \tag{12}$$

The total loss to distill an on-device model is:

$$\mathcal{L}_{\text{DistillGaze}} = \sum_{\boldsymbol{X}\in\mathcal{D}_{\text{real}}} \left[\lambda_t(\mathcal{L}_{\text{KD}} + \mathcal{L}_{\text{Pseudo-}t}) + (1-\lambda_t)(\mathcal{L}_{\text{SD}}^{\text{On-device}} + \mathcal{L}_{\text{Pseudo-}e})\right], \tag{13}$$

where λ_t is a hyperparameter that weights the two sources of supervision: the optimized VFM teacher and the EMA student. A cosine schedule controls their relative contribution: it emphasizes distillation from the optimized VFM early in training and gradually shifts toward EMA self-distillation. This allows the student to first inherit structured representations from the optimized VFM and then refine them using stable EMA targets. At inference time, we deploy only the on-device student network f_s.

4 Experimental design

4.1 Dataset

We evaluate our method on data collected with Project Aria [29]. Following the Aria convention, our real dataset comprises 6,299 recordings from 2,222 crowd-sourced participants, partitioned into training and validation splits of 1,825 and 397 participants, respectively. To simulate the unlabeled-data setting, we withhold the ground-truth gaze labels for the training set. Input images are captured by two global-shutter monochrome cameras under diffused infrared illumination. The ET cameras record at 20 fps for approximately 60 seconds per recording; we temporally subsample by a factor of 10 for training efficiency. Each eye camera captures at 640×480 resolution, which we downsample to 320×240 following [29]. The dataset provides aligned 3D gaze targets. We compute gaze rays using a fixed origin derived from a nominal interpupillary distance of 63.5 mm. Our synthetic data comprise 165K frames from 998 subjects, rendered in Blender.

4.2 Implementation Details

Our on-device ET pipeline contains 256K parameters and uses an FBNet [44] backbone shared between images of the left and right eyes. We train with AdamW using a cosine learning-rate schedule with a 10% warm-up period and a batch size of 256. The learning rate is adjusted between 10^{-3} and 10^{-5}, and each experiment trains for up to 50,000 iterations. For self-distillation, the EMA teacher weights are updated every 100 iterations with the momentum adjusted between 0.95 and 0.99. The feature projection head consists of three fully connected layers with GELU activations. For VFM experiments, we initialize from the ViT-B variant of DINOv3 [38] (86M parameters). All experiments were conducted on 4× NVIDIA A100 80GB GPUs.

Image transformations. We employ two augmentation pipelines: strong augmentation for the student, and weak augmentation for the teacher and EMA student. The weak pipeline consists of gamma jitter and random scaling. The strong pipeline additionally applies camera and illumination augmentations, each with probability 0.3. Camera augmentations simulate sensor degradation through modulation transfer function (MTF) filtering, motion blur, and compression artifacts, while illumination augmentations comprise brightness and contrast adjustment, glint inpainting, random shadows, and coarse dropout.

4.3 Evaluation Protocol

We measure gaze estimation accuracy as the angular error (in degrees) between the predicted and ground-truth 3D gaze vectors, averaged over the left and right eyes for each frame. Following [2,29], we report results using an Error-User (EU) table to capture both frame-level and subject-level performance. We summarize each user's per-frame error distribution via percentiles E50, E75, and E90, then aggregate across users at U50, U75, and U90. Thus, E50U50 reflects the median error of the median user, whereas E90U90 characterizes tail performance. For efficiency, we uniformly sample 64 frames per recording: 9 frames are reserved for test-time personalization, and metrics are computed on the remaining 55. We report 95% confidence intervals via a hierarchical bootstrap over users and frames (1,000 iterations).

5 Results

We first evaluate DINOv3 VFMs in Sec. 5.1, then present our optimized VFM performance in Sec. 5.2, the on-device distillation in Sec. 5.3, and ablation studies for our training dynamics in Sec. 5.4.

5.1 Performance of Off-the-Shelf VFMs on Eye Tracking

Given the strong transfer performance of VFMs across diverse vision tasks, we first evaluate whether off-the-shelf representations directly benefit synthetic supervised gaze estimation. We extract embeddings from the ViT-B variant of

Table 1: Gaze estimation performance comparing linear-evaluated foundation models and models optimized with synthetic and unlabeled real data. 95% confidence intervals in parentheses. Inf. params: number of parameters at inference, SynSup: synthetic supervised, SynFT: synthetic fine-tuning.

	Inf. params	E50U50 ↓	E75U75 ↓	E90U90 ↓
On-device SynSup ♦	256 K	3.48 (0.19)	7.41 (0.53)	14.84 (2.03)
DINOv3 [38] Linear Probe	86 M	5.47 (0.23)	10.64 (0.61)	18.74 (1.60)
DINOv3 [38] SynFT	86 M	2.01 (0.11)	4.29 (0.46)	12.11 (1.50)
DARE-GRAM [30]	86 M	4.84 (0.24)	9.36 (0.54)	17.89 (1.65)
Optimized VFM ▲	86 M	**1.33 (0.07)**	**3.07 (0.34)**	**8.32 (2.10)**

the DINOv3 family [38], fit a linear regressor on synthetic data, and apply our per-subject calibration protocol (Sec. 4.3).

As shown in the first section of Tab. 1, DINOv3 with linear probe trained on synthetic data yields substantially lower accuracy than our small on-device model trained with synthetic supervision (on-device SynSup), despite having 336× more parameters. This observation is consistent across all DINOv3 variants, spanning both ViT and ConvNeXt architectures, as illustrated in Fig. 1b. Fine-tuning DINOv3 ViT-B on synthetic data (DINOv3 SynFT) substantially improves performance (Tab. 1), reducing E50U50 from 5.47° to 2.01° and surpassing the on-device baseline by leveraging the VFM prior and larger model capacity. This confirms our hypothesis: VFMs provide a useful initialization but require further optimization on domain-relevant data for ET.

5.2 Optimize VFM Teacher for Eye Tracking

We next study how to adapt the VFM teacher for eye tracking using synthetic labeled data and unlabeled real data. Unsupervised domain adaptation (UDA) is a natural way to leverage synthetic data for VFM optimization. DARE-GRAM [30] is a state-of-the-art UDA method designed for regression tasks. However, as shown in Tab. 1, it performed worse than DINOv3 SynFT, potentially due to overfitting to the synthetic data distribution, which may amplify the mismatch between synthetic and real data.

In contrast, we optimize the VFM with the combined self-distillation and synthetic supervision objective described in Sec. 3.2. The result is shown in Tab. 1. Compared to DINOv3 SynFT, our optimized VFM reduces E50U50 from 2.01° to 1.33° (33.8% improvement). The gains are consistent in the tail: E90U90 decreases from 12.11° to 8.32° (31.3% improvement). In particular, our optimized VFM substantially outperforms the synthetic on-device baseline (1.33° vs. 3.48° at E50U50, a 61.8% reduction). This shows that self-supervised learning on unlabeled real data can effectively bridge the synthetic-to-real gap. Our optimized ViT-B VFM serves as the distillation source in Sec. 5.3, since our ultimate goal is a deployable on-device model.

Table 2: On-device model performance comparison. 95% confidence intervals in parentheses. Fully supervised upper bound is provided as reference, as it is trained on data that would not be available at the time of product development (bold = best among on-device students). SynSup: synthetic supervised, Inf. params: number of parameters at inference time.

	Inf. params	E50U50 ↓	E75U75 ↓	E90U90 ↓
Methods without VFM				
On-device SynSup ♦	256 K	3.48 (0.19)	7.41 (0.53)	14.84 (2.03)
DARE-GRAM [30]	256 K	2.96 (0.17)	6.20 (0.47)	13.41 (1.73)
Self-distillation	256 K	2.20 (0.11)	4.64 (0.46)	10.95 (2.10)
Distillation with Optimized VFM				
Optimized VFM ▲	86 M	1.33 (0.07)	3.07 (0.34)	8.32 (2.10)
Pseudo labels	256 K	1.59 (0.09)	3.49 (0.30)	**8.19 (1.56)**
SP [40]	256 K	1.59 (0.09)	3.50 (0.30)	8.21 (1.71)
VIC-KD [15]	256 K	1.46 (0.08)	3.32 (0.35)	8.40 (2.01)
Community KD [24]	256 K	1.48 (0.09)	**3.24 (0.31)**	8.47 (1.88)
DistillGaze (ours) ★	256 K	**1.44 (0.09)**	3.29 (0.32)	8.45 (2.04)
Fully supervised (upper bound)	256 K	0.82 (0.05)	1.94 (0.26)	5.91 (1.61)

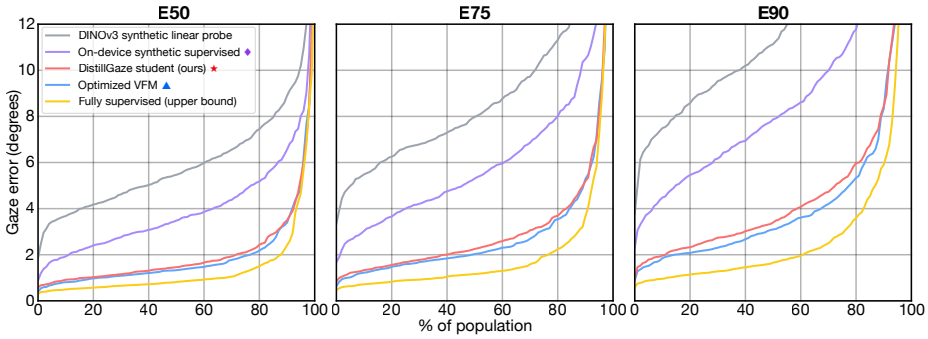


Fig. 3: Population coverage across error percentiles. We plot the cumulative distribution of users at E50, E75, and E90. Our optimized VFM and DistillGaze student exhibit closely matched performance, demonstrating effective knowledge transfer. Both methods approach the fully supervised on-device upper bound (yellow).

5.3 Distill to On-Device Model

We evaluate our distillation framework (Sec. 3.3) to transfer knowledge from our optimized VFM to an on-device ET model. We compare DistillGaze against DARE-GRAM [30], a state-of-the-art UDA method for adaptation without teacher, and several knowledge distillation approaches utilizing the teacher, including VIC-KD [15] and Community KD [24]. The results are summarized in Tab. 2.

Methods without VFM. To demonstrate the value of optimizing the VFM, we trained the small, on-device model without VFM teacher, using synthetic supervision and unlabeled real data. We conducted two experiments. First, we trained

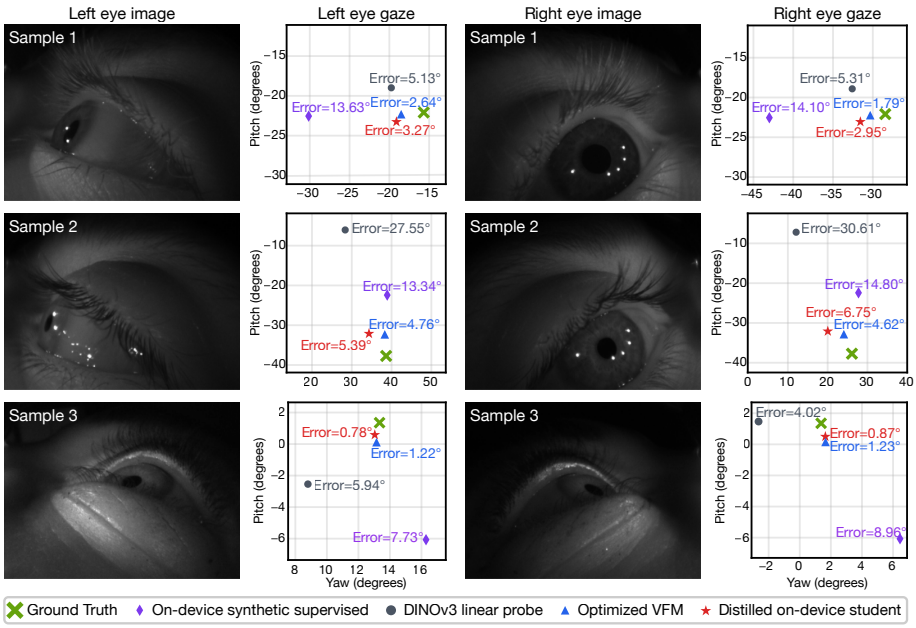


Fig. 4: Qualitative comparisons of gaze predictions. Pitch-yaw plots (degrees) compare ground truth (green) against four methods. The on-device synthetic supervised model (purple) and the frozen DINOv3 linear probe (gray) yield large errors, highlighting the need for adaptation beyond synthetic-only training or frozen VFM features. Our distilled on-device student (red) achieves accuracy comparable to the optimized VFM teacher (blue) while requiring 336× fewer parameters.

DARE-GRAM, a state-of-the-art UDA method. This reduced E50U50 to 2.96°, outperforming training with synthetic supervision. Second, we conducted self-distillation following Sec. 3.2, using an on-device model, which reduces E50U50 to 2.20°, a 36.8% improvement. This shows that training with unlabeled real data significantly improves model performance even without a VFM teacher, but the remaining gap to the optimized VFM highlights the importance of the VFM prior and learning capacity.

Distillation with Optimized VFM. The methods distilled from our optimized VFM yield substantial improvements over the baselines without VFM. DistillGaze achieves the best E50U50 performance, which reflects the typical user experience during normal conditions. Community KD has the best performance on E75U75, and training with pseudo labels only achieves the best E90U90. We note that tail samples in crowd-sourced datasets are more susceptible to label noise, arising from ambiguous gaze targets or inconsistent collection conditions, which may disproportionately affect tail metrics. In this regime, directly regressing to predictions from a strong optimized teacher is already a robust objective, which limits the marginal gains from additional distillation losses. With an optimized strong teacher, student performance may be driven more by teacher quality than

Table 3: Ablations: teacher loss function. 95% confidence intervals in parentheses.

	E50U50 ↓	E75U75 ↓	E90U90 ↓
Optimized VFM ▲	**1.33 (0.07)**	**3.07 (0.34)**	**8.32 (2.10)**
Optimized VFM ▲ using DINO loss	1.50 (0.10)	3.32 (0.38)	8.86 (1.94)

the specific distillation objective. In scenarios where a strong teacher is not available, our proposed method provides more robust performance gains than these baselines, as shown in Supplementary Sec. B.4 with a ConvNeXt-S teacher.

Compared to the baseline model, on-device synthetic supervision, DistillGaze reduces E50U50 from 3.48° to 1.44° (58.6% improvement). Performance improvements also extend to the tail: E90U90 decreases from 14.84° to 8.45° (43.1% reduction). We show population coverage charts at error percentiles E50, E75, and E90 in Fig. 3, which visualize the complete distribution between users. Our optimized VFM and DistillGaze student exhibit closely matched performance across all user percentiles at E50 and E75, with a modest degradation at E90, demonstrating strong knowledge transfer during distillation. Both methods show a strong performance approaching a fully supervised on-device model trained with labeled data that would not be available during product development (last row in Tab. 2). Qualitative comparisons are presented in Fig. 4.

5.4 Ablation Studies

Teacher loss function. Many conventional self-supervised methods employ centering and prototype-based soft clustering to prevent representation collapse [7, 32]. However, we find that a simple MSE regression loss substantially outperforms the DINO objective in our setting (Tab. 3). We attribute this to two key factors. First, soft clustering encourages discrete, categorical representations that may limit the fine-grained resolution required for continuous regression tasks like gaze estimation. Second, our use of synthetic supervision provides an inherent regularization signal that prevents collapse, reducing the reliance on these explicit mechanisms employed in self-supervised settings.

VFM model architecture. One of our hypotheses was that ConvNeXt mod-

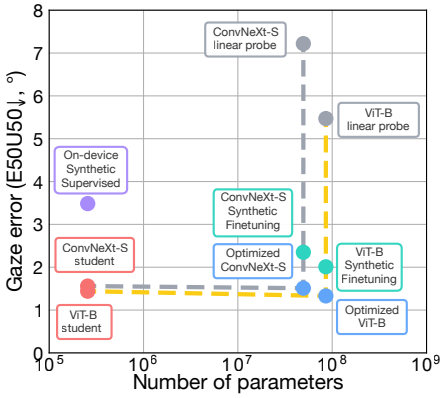


Fig. 5: Ablations on VFM architecture. Gaze error (E50U50 ↓, °) vs. parameter count for ConvNeXt-S and ViT-B backbones. Linear probing (gray) performs poorly despite high parameter counts. Synthetic fine-tuning and our optimization substantially reduce error. Distilled on-device students (left) approach optimized VFM accuracy with orders of magnitude fewer parameters.

els would perform better than ViTs, since eye tracking features are concentrated near the pupil, and model performance may be constrained by ViT tokenization resolution. Interestingly, as we can observe in Fig. 5, this is not the case. ViT-B exhibits stronger linear probe performance, and this advantage propagates through synthetic fine-tuning and self-distillation. However, the performance gap progressively narrows. In particular, both VFMs substantially outperform the on-device model with synthetic supervision baseline, which we attribute to the larger capacity of the models and the strong priors inherited from DINOv3 initialization. We speculate that peripheral regions of the eye visible within each patch may provide sufficient contextual cues, enabling ViT to match or exceed ConvNeXt's performance.

5.5 Limitations and Future Work

Although our method substantially improves over on-device baselines, a gap remains to the fully supervised upper bound, which uses labeled data that would not be available during product development. This gap is larger in the tail, suggesting that challenging cases, such as users with atypical eye appearance or extreme gaze angles, remain difficult without labeled supervision.

Several promising directions may help narrow this gap, including bridging the domain discrepancy between real and synthetic data through more advanced generative modeling, designing stronger self-supervised objectives, or incorporating lightweight online adaptation at inference time. In addition, all training and evaluation in this present work are conducted on a single device configuration. Extending the proposed framework to multiple devices with varying camera geometries, illumination, and sensor characteristics is an important direction for future work.

6 Conclusion

We presented DistillGaze, a framework for the rapid deployment of on-device eye tracking systems that leverages visual foundation models without requiring real ground-truth gaze supervision. Our analysis reveals that while DINOv3 representations cluster by subject identity, intra-subject distributions exhibit smooth gaze-correlated structure amenable to adaptation. We exploit this finding through a two-stage approach: first, optimizing DINOv3 via self-distillation combined with synthetic supervision and unlabeled real data, and then distilling the result into a compact on-device model. The final student achieves 58.6% and 43.1% reductions in E50U50 and E90U90, respectively, over a synthetic-supervised baseline, without any real ground-truth labels and with no increase in model size. By eliminating the need for labeled real data, DistillGaze enables rapid adaptation of the model to new hardware configurations, reducing the development cycle from months to days.

References

1. Albuquerque, I., Naik, N., Li, J., Keskar, N., Socher, R.: Improving out-of-distribution generalization via multi-task self-supervised pretraining. arXiv preprint arXiv:2003.13525 (2020) 4
2. Aziz, S., Lohr, D.J., Friedman, L., Komogortsev, O.: Evaluation of eye tracking signal quality for virtual reality applications: A case study in the meta quest pro. In: Proceedings of the 2024 Symposium on Eye Tracking Research and Applications. pp. 1–8 (2024) 10
3. Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., et al.: A cookbook of self-supervised learning. arXiv preprint arXiv:2304.12210 (2023) 5
4. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021) 5
5. Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906 (2021) 5, 9
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025) 3, 5
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: IEEE International Conference on Computer Vision (ICCV). pp. 9650–9660 (2021) 5, 8, 14
8. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024) 3
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML). pp. 1597–1607. PmLR (2020) 5
10. Cheng, Y., Bao, Y., Lu, F.: Puregaze: Purifying gaze feature for generalizable gaze estimation. In: Conference on Artificial Intelligence (AAAI). vol. 36, pp. 436–443 (2022) 4
11. Cheng, Y., Lu, F.: Gaze estimation using transformer. In: International Conference on Pattern Recognition (ICPR). pp. 3341–3347. IEEE (2022) 4
12. Fuhl, W., Kasneci, G., Kasneci, E.: Teyed: Over 20 million real-world eye images with pupil, eyelid, and iris 2d and 3d segmentations, 2d and 3d landmarks, 3d eyeball, gaze vector, and eye movement types. arXiv preprint arXiv:2102.02115 (2021) 4
13. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems (NIPS) **33**, 21271–21284 (2020) 5, 8
14. Guestrin, E.D., Eizenman, M.: General theory of remote gaze estimation using the pupil center and corneal reflections. IEEE Transactions on biomedical engineering **53**(6), 1124–1133 (2006) 4
15. Guimarães, H.R., Pimentel, A., Avila, A., Falk, T.H.: Vic-kd: Variance-invariance-covariance knowledge distillation to make keyword spotting more robust against adversarial attacks. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 12196–12200. IEEE (2024) 8, 12

16. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (2022) 5
17. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9729–9738 (2020) 5
18. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 5
19. Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Yokoya, N., Li, H., Ghamisi, P., Jia, X., et al.: Spectralgpt: Spectral remote sensing foundation model. IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI) **46**(8), 5227–5244 (2024) 3
20. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: IEEE International Conference on Computer Vision (ICCV). pp. 6912–6921 (2019) 4
21. Khirodkar, R., Wen, H., Martinez, J., Dong, Y., Su, Z., Saito, S.: Sapiens2. In: The Fourteenth International Conference on Learning Representations (2026) 5
22. Kim, J., Stengel, M., Majercik, A., De Mello, S., Dunn, D., Laine, S., McGuire, M., Luebke, D.: Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation. In: Proceedings of the 2019 CHI conference on human factors in computing systems. pp. 1–12 (2019) 4
23. Kondepudi, A., Pekmezci, M., Hou, X., Scotford, K., Jiang, C., Rao, A., Harake, E.S., Chowdury, A., Al-Holou, W., Wang, L., et al.: Foundation models for fast, label-free detection of glioma infiltration. Nature **637**(8045), 439–445 (2025) 3
24. Lee, H., Hou, R., Kim, J., Liang, D., Zhang, H., Hwang, S., Min, A.: Co-training and co-distillation for quality improvement and compression of language models. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 7458–7467 (2023) 5, 8, 12
25. Li, G., Meka, A., Mueller, F., Buehler, M.C., Hilliges, O., Beeler, T.: Eyenerf: a hybrid representation for photorealistic synthesis, animation and relighting of human eyes. ACM Trans. on Graphics (TOG) **41**(4), 1–16 (2022) 4
26. Lin, E.Y., Ding, Y., Kundu, J., An, Y., El-Haddad, M.T., Fix, A.: Digitally prototype your eye tracker: Simulating hardware performance using 3d synthetic data. arXiv preprint arXiv:2503.16742 (2025) 1, 4
27. Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J.: Remoteclip: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024) 3
28. Liu, M., Li, Y., Liu, H.: 3d gaze estimation for head-mounted eye tracking system with auto-calibration method. IEEE Access **8**, 104207–104215 (2020) 4
29. Mansour, Y., Fernandes, A.S., Somasundaram, K., Hefny, T., Shakeri, M., Komogortsev, O., Sharma, A., Proulx, M.J.: Enabling eye tracking for crowd-sourced data collection with project aria. IEEE Access (2025) 1, 2, 3, 4, 6, 7, 9, 10
30. Nejjar, I., Wang, Q., Fink, O.: Dare-gram: Unsupervised domain adaptation regression by aligning inverse gram matrices. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11744–11754 (2023) 11, 12
31. Nguyen, V.D., Bailey, R., Diaz, G.J., Ma, C., Fix, A., Ororbia, A.: Deep domain adaptation: A sim2real neural approach for improving eye-tracking systems. Proceedings of the ACM on Computer Graphics and Interactive Techniques **7**(2), 1–17 (2024) 4

32. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 5, 14
33. Palmero, C., Sharma, A., Behrendt, K., Krishnakumar, K., Komogortsev, O.V., Talathi, S.S.: Openeds2020: Open eyes dataset. arXiv preprint arXiv:2005.03876 (2020) 4
34. Porta, S., Bossavit, B., Cabeza, R., Larumbe-Bergera, A., Garde, G., Villanueva, A.: U2eyes: A binocular dataset for eye tracking and gaze estimation. In: IEEE International Conference on Computer Vision (ICCV) Workshop (2019) 4
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PmLR (2021) 5
36. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets (2015) 5
37. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2107–2116 (2017) 4
38. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025) 3, 5, 10, 11
39. Tian, Y., Krishnan, D., Isola, P.: Contrastive Representation Distillation. In: International Conference on Learning Representations (2020) 5
40. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: IEEE International Conference on Computer Vision (ICCV). pp. 1365–1374 (2019) 5, 12
41. Wang, S., Yu, Z., Jiang, X., Lan, S., Shi, M., Chang, N., Kautz, J., Li, Y., Alvarez, J.M.: Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22442–22452 (2025) 3
42. Wood, E., Baltrušaitis, T., Morency, L.P., Robinson, P., Bulling, A.: Learning an appearance-based gaze estimator from one million synthesised images. In: Proceedings of the ninth biennial ACM symposium on eye tracking research & applications. pp. 131–138 (2016) 4
43. Wood, E., Baltrusaitis, T., Zhang, X., Sugano, Y., Robinson, P., Bulling, A.: Rendering of eyes for eye-shape registration and gaze estimation. In: IEEE International Conference on Computer Vision (ICCV). pp. 3756–3764 (2015) 4
44. Wu, B., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., Tian, Y., Vajda, P., Jia, Y., Keutzer, K.: Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10734–10742 (2019) 10
45. Xiao, Y., Bai, X., Chen, B., Su, H., He, H., Xie, L., Yin, E.: De^ 2gaze: Deformable and decoupled representation learning for 3d gaze estimation. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3091–3100 (2025) 4
46. Xu, J., Xiao, L., López, A.M.: Self-supervised domain adaptation for computer vision tasks. IEEE Access **7**, 156694–156706 (2019) 4
47. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016) 5

48. Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O.: Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In: European Conference on Computer Vision (ECCV). pp. 365–381 (2020) 4
49. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4511–4520 (2015) 4
50. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021) 5
51. Zhu, Z., Ji, Q.: Eye gaze tracking under natural head movements. In: Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. 918–923. IEEE (2005) 4