SG-GAZE: STRUCTURALLY AND GEOMETRICALLY CONSISTENT REPRESENTATION LEARNING FOR GENERALIZABLE 3D GAZE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning accurate and generalizable 3D gaze representations remains challenging due to the lack of a unified and physically meaningful representation. Existing methods rely on either appearance features or simplified geometric modeling, but fail to jointly capture geometric and structural consistency. They exhibit poor cross-domain generalization and typically require large-scale multiview datasets to mitigate viewpoint variation, yet still struggle with domain shifts between controlled and in-the-wild settings. To address these issues, we propose **SG-Gaze**, a dual-branch framework that learns a **S**tructurally and **G**eometrically Consistent **R**epresentation **(SGR)** for gaze estimation. The analytical branch embeds features into a geodesically aligned spherical manifold for interpretable regression, while the model-guided branch reconstructs 3D eyeball structure under weak 2D edge supervision. Through adversarial training, the resulting SGR is simultaneously appearance discriminative, structurally faithful, and geometrically consistent. To further improve robustness, we introduce View-Consistent Regularization, which augments training SGR with synthetic view perturbations and enforces rotation-equivariant consistency across gaze vectors and structural projections. This reduces reliance on costly multiview data and narrows cross-domain gaps. Extensive experiments on synthetic and real-world datasets show that SG-Gaze achieves state-of-the-art accuracy and strong cross-domain generalization in 12 challenging transfer scenarios. Our work demonstrates the importance of unifying structurally and geometrically consistent representation with equivariant regularization, providing broader insight into building more interpretable and generalizable models.

1 INTRODUCTION

Eye gaze estimation has broad applications in human–computer interaction (Admoni & Scassellati, 2017; Terzioğlu et al., 2020), driver monitoring (Cañas et al., 2025; Cheng et al., 2024), and immersive AR/VR systems (Burova et al., 2020; Konrad et al., 2020). Beyond these applications, gaze estimation also provides a valuable testbed for studying fundamental challenges in representation learning. Accurate and generalizable 3D gaze representation is a challenging problem that reflects broader questions in representation learning: how to construct interpretable, physically meaningful features that generalize across viewpoints, devices, and domains. The core challenge lies in learning a unified representation that captures both appearance cues and geometric structure.

Despite rapid advances, existing methods still face two key limitations: (1) **Lack of unified physical representation.** As shown in Fig. 1, appearance-based methods (Zhang et al., 2015; Chen & Shi, 2018; Cheng et al., 2020b) directly regress gaze from image features without explicit geometric constraints (Fig. 1(a)), while model-based methods (Chen et al., 2008; Hennessey et al., 2006; Świrski & Dodgson, 2013) reconstruct eyeball geometry but require personal calibration and dedicated devices (Fig. 1(b)). These two isolated methods fail to combine geometric constraints with structural modeling to form a unified representation, resulting in weak interpretability and poor generalization across subjects and devices. (2) **Limited generalization across viewpoints and domains.** Gaze estimation models must be robust to diverse head poses, camera viewpoints, and environmental variations. Existing models are sensitive to such changes. To compensate, they rely on large-scale multiview training data, but are still insufficient to bridge controlled and in-the-wild domains gaps (Fig. 1(f)), because training with common gaze datasets often results in poor cross-dataset performance. This limitation stems from the lack of a representation that enforces rotational equivariance and geometric-structural consistency, which is critical for generalizable gaze modeling.

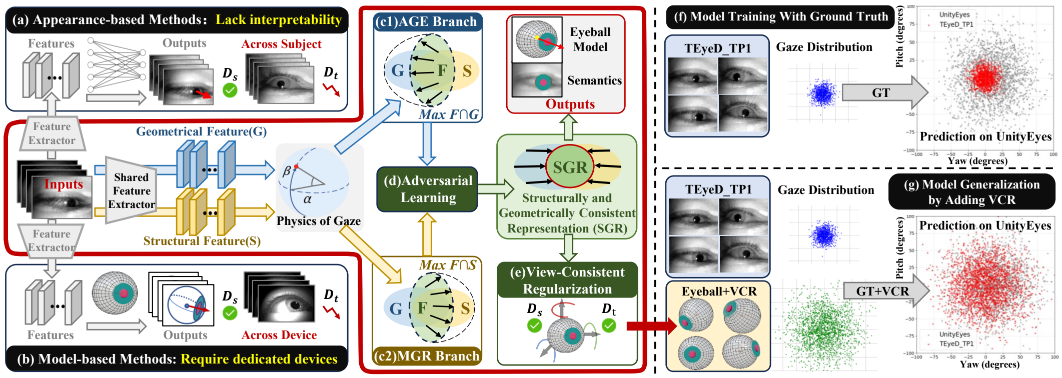

Figure 1: Overview of gaze representation learning methods. (a) Appearance-based methods lack geometric constraints and interpretability. (b) Model-based methods require dedicated devices. (c) Our SG-Gaze unifies both methods by learning Geometric**(G)** and Structural**(S)** features grounded in gaze physics. (d) Adversarial learning fusing Max$(F \cap G)$ and Max$(F \cap S)$ into a unified Structurally and Geometrically Consistent Representation (SGR). (e-g) VCR enforces rotation-equivariant consistency to cover a larger field-of-view to mitigate domain gaps compared to GT supervision alone.

To address these challenges, we propose **SG-Gaze**, a dual-branch framework designed to learn a structurally and geometrically consistent representation **(SGR)** for gaze estimation. Unlike traditional methods that focus solely on appearance or model, SGR unifies physical priors with representation learning to achieve both interpretability and generalization. Given a sequence of near-eye images, a shared backbone extracts discriminative visual features, preserving appearance cues necessary for 3D gaze modeling. These features are then processed by two cooperative branches. The **A**nalytical **G**aze **E**stimation **(AGE)** branch encodes features on a geodesically aligned spherical manifold (Fig. 1(c1)), capturing inherent physical constraints for interpretable regression. In parallel, the **M**odel-**G**uided **R**econstruction **(MGR)** branch reconstructs 3D eyeball structures under weak 2D edge supervision (Fig. 1(c2)), enforcing structural fidelity and anatomical consistency. A dedicated adversarial alignment module bridges the two branches (Fig. 1(d)), producing SGR, which unifies appearance discriminative features with structurally and geometrically faithful representations. Furthermore, to improve generalization across unseen viewpoints and domains, we introduce **V**iew-**C**onsistent **R**egularization **(VCR)**, which augments training SGR with synthetic viewpoint perturbations and enforces rotation-equivariant consistency of gaze vectors and structural projections (Fig. 1(e)). This strategy effectively simulates multi-view conditions without costly data collection and mitigates controlled-to-real domain gaps by covering a larger field-of-view (Fig. 1(g)).

Finally, SG-Gaze jointly predicts 3D eyeball, gaze direction and auxiliary 2D semantic projections, achieving interpretable and generalizable representations. Comprehensive experiments on synthetic dataset UnityEyes (Wood et al., 2016) and real-world datasets TEyeD (Fuhl et al., 2021), LPW (Tonsen et al., 2016) show that SG-Gaze achieves state-of-the-art accuracy and strong cross-dataset generalization, validating the benefits of learning a unified structurally and geometrically consistent representation with equivariant regularization. To summarize, our main contributions are three-fold:

- We propose SG-Gaze, a unified dual-branch framework. By combining AGE and MGR based on the physical principles of gaze, SG-Gaze learns a Structurally and Geometrically Consistent Representation (SGR), which is interpretable and physically meaningful.
- We introduce View-Consistent Regularization (VCR), a training strategy that enforces rotation-equivariant consistency of gaze vector and structure projection under synthetic viewpoint perturbations, alleviating domain gaps in a principled way.
- Experimental results show that SG-Gaze achieves state-of-the-art accuracy and strong generalization, improving the baseline by up to 38.61% without touching target domain data. The approach achieves consistent improvements across 12 challenging transfer scenarios.

2 Motivation: Structural and Geometric Consistency

Our representation of gaze is grounded in two fundamental physical properties of human vision:
Fact 1: Gaze vectors lie on a unit 3D sphere Gaze direction can be defined as a unit vector from the eyeball center to a fixation point. Hence, all gaze vectors reside on the surface of a unit sphere

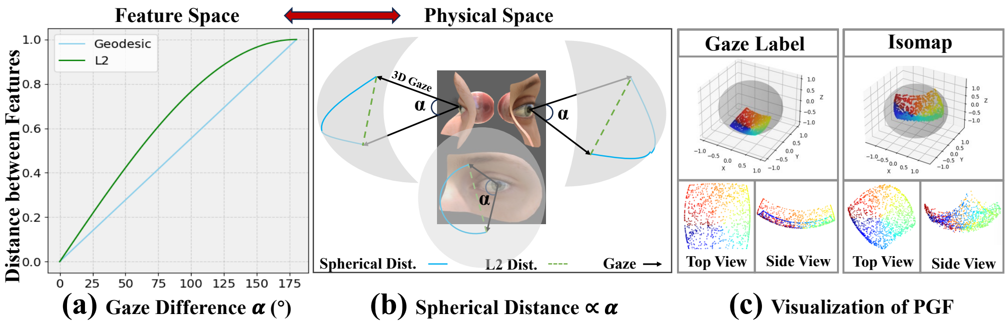

Figure 2: Analysis of gaze features in relation to physical geometry. (a) L2 and geodesic distances vs. angular differences on UnityEyes. (b) Geodesic distance on feature sphere proportional to gaze differences. (c) Visualization of the principal gaze feature after projection using geodesic distance, where features exhibit a spherical distribution consistent with gaze labels.

$\mathbb{S}^2$. As shown in Fig. 2(b), the geodesic distance between two points on this sphere corresponds to the angular difference between gaze vectors, defining a structured manifold for gaze representation.
Fact 2: The eyeball is a physical sphere Anatomically, the eyeball approximates a rigid 3D sphere, where the gaze direction is determined by the geometric relation between the eyeball center and the pupil center. This means that gaze estimation does not simply rely on image pixels or implicit features but is a geometric problem based on the eye structure and subject to physical constraints.
Spherical Structure in Feature Space To examine wether feature space preserves geometric structure of gaze, we conduct experiments on UnityEyes (Wood et al., 2016). Source domain is denoted as $D_s = \{x_i, y_i\}_{i=1}^N$, where x_i is eye image and y_i the corresponding gaze vector. We pretrain a ResNet-18 (He et al., 2016) feature extractor $F_{\theta_1}(\cdot)$ with L_1 loss: $f_i = F_{\theta_1}(x_i)$, $\min_{\theta_1,\theta_2} \sum_{i=1}^N L_1\left(y_i, L_{\theta_2}(f_i)\right)$, where $f_i \in \mathbb{R}^{512}$ and $L_{\theta_2}(\cdot)$ is the final fully connected layer. We compute the geodesic distance between feature embeddings and compare it with angular difference of their ground-truth gaze vectors. **Observation**: The two exhibit a strong linear correlation, indicating that the feature space encodes the spherical topology of gaze directions. Motivated by this, we follow AGG (Bao & Lu, 2024) and extract a low-dimensional *Principal Gaze Feature (PGF)* by applying Isometric Mapping (Isomap) (Tenenbaum et al., 2000) to project f_i into a 3D subspace. As shown in Fig. 2(c), the PGFs lie approximately on the surface of a 3D sphere, preserving the structural geometry of gaze.

These physical principles motivate reformulating gaze estimation as a feature learning problem under structural and geometric consistency. We advocate learning a unified and physically meaningful representation that is interpretable and rotation-consistent instead of regressing gaze vectors directly. We refer this as the **Structurally and Geometrically Consistent Representation (SGR)** of gaze.

3 Methodology

We define the Structurally and Geometrically Consistent Gaze Representation $\mathrm{SGR} \in \mathbb{R}^d$ that satisfies three key physical constraints. **(1) Geometrical consistency**: its 3D projection must lie on the unit sphere, $\|\mathrm{SGR}_{\mathrm{proj}}\| = 1$, thereby preserving geodesic locality in $\mathbb{S}^2$ and reflecting the angular structure of gaze space. **(2) Structural decodability**: SGR should embed geometric components that allow the recovery of eyeball parameters via a decoder, $E_{para} = D(\mathrm{SGR})$, ensuring interpretability through model reconstruction. **(3) Rotation equivariance**: the decoded structure must be consistent under 3D rotations, i.e., $F(T_P(\mathrm{SGR})) \approx P \cdot D(\mathrm{SGR})$, where T_P is a learned rotation operator and P a physical rotation matrix. Together, these constraints enforce that SGR not only predicts gaze direction but also aligns with anatomical geometry and remains robust under geometric transformations, forming a unified and physically grounded representation for gaze estimation.

3.1 Overview

The overall pipeline is illustrated in Fig. 3. Given a sequence of near-eye image frames $I = \{I_1, \ldots, I_n\}$, our framework SG-Gaze estimates the 3D eyeball structure e^m, gaze direction e^g, and 2D semantics e^s through a unified function Φ, i.e., $\{e^m, e^g, e^s\} = \Phi(I)$. The architecture consists of three modules: (1) The input frames I are encoded by a backbone B (ResNet-18) into visual features $F = B(I)$ [Fig. 3(a)]. (2) The decoder Φ_{dec} contains two cooperative branches: the Analytical Gaze Estimation **(AGE)** branch projects features onto a geodesically aligned spherical manifold to yield an interpretable embedding for gaze regression[Fig. 3(b)]; the Model-Guided Reconstruction **(MGR)**

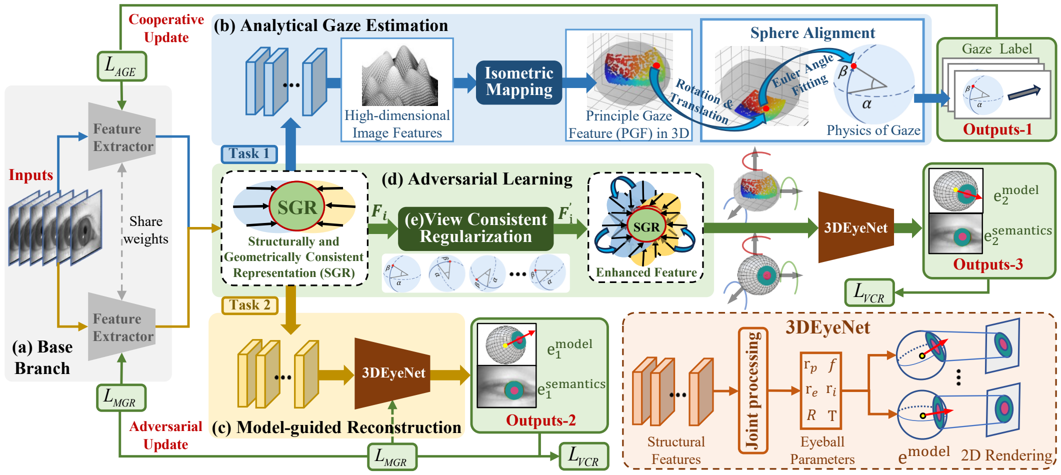

Figure 3: Overview of SG-Gaze. Given a sequence of near-eye frames I, the framework Φ extracts features through a shared backbone and processes them with AGE branch and MGR branch, enhanced by adversarial learning and view consistent regularization. SG-Gaze jointly predicts a 3D eyeball model e^m, 3D gaze e^g and 2D semantics e^s, formalized as $\{e^m, e^g, e^s\} = \Phi(I)$.

branch predicts the eyeball structure (iris, pupil, sphere center) and jointly optimizes the gaze via physical modeling constraints through 3DEyeNet[Fig. 3(c)]. Both branches share intermediate representation and are jointly optimized. (3) We introduce a cross-task adversarial discriminator in Φ_{afr} [Fig. 3(d)], encouraging AGE and MGR to produce structurally and geometrically consistent features. In addition, View-Consistent Regularization **(VCR)** applies synthetic rotations and enforces rotation-equivariant consistency of SGR across gaze vectors and structural projections, improving robustness to viewpoint variation [Fig. 3(e)]. Through the joint optimization of AGE, MGR and VCR, SG-Gaze learns a Structurally and Geometrically Consistent Representation (SGR) that unifies appearance cues and physical modeling. It outputs the optimized 3D eyeball model e^m, gaze vector e^g, and auxiliary semantic projections e^s with improved accuracy and interpretability.

3.2 Analytical Gaze Estimation

The Analytical Gaze Estimation (AGE) branch predicts 3D gaze from shared features $\mathbf{f}_i$ via a geometrically interpretable process. Unlike black-box regression, AGE leverages spherical projection and analytical fitting to construct gaze vectors based on their angular position on a unit sphere.
Spherical Projection As observed in Section 2, feature geodesic distances are linearly correlated with gaze angular differences. To preserve this topology, we apply Isomap to embed features $\mathbf{f}_i \in \mathbb{R}^{512}$ into a 3D space: $\{\mathbf{p}_i\}_{i=1}^{N'} = \text{Isomap}(\{f_i\}_{i=1}^{N'})$, where $p_i \in \mathbb{R}^3$ represents the Principal Gaze Feature (PGF). The resulting embedded points approximately lie on the surface of a 3D sphere, reflecting gaze-aware topology.

Spherical Fitting We adopt a Spherical Fitting (SF) algorithm to map projected features to gaze angles analytically. Following the previous work (Bao & Lu, 2024), the sphere center O_c is estimated, followed by a rotation R: $p'_i = R(p_i - O_c) = (x'_i, y'_i, z'_i)^T$. Euler angles (θ'_i, ψ'_i) are then derived from the rotated point p'_i via simple linear projections: $\theta'_i = k_1 \arctan\left(\frac{x'_i}{z'_i}\right) + b_1,\ \psi'_i = k_2 \arcsin(y'_i) + b_2$, where (k_1, k_2, b_1, b_2) are learnable scaling and biases parameters. The final 3D gaze vector is recovered as $y'_i = \text{SF}_{\theta_s}(p_i),\ \theta_s = \{O_c, R, k_1, k_2, b_1, b_2\}$. Parameters θ_s are optimized by minimizing angular error: $\min_{\theta_s} \sum_{i=1}^{N} \text{Angular}(y_i, \text{SF}_{\theta_s}(p_i))$. After spherical mapping, AGE enforces gaze-aware geometry, stabilizes learning and provides a basis for structure-aware fusion with the MGR branch.

3.3 Model-Guided Reconstruction

While AGE enforces geometric alignment, it ignores anatomical constraints. MGR branch addresses this by introducing a parametric 3D eyeball, ensuring gaze prediction is physically interpretable.

Canonical Eyeball Modeling As shown in Fig. 4(a). Consistent with common practice in anatomical eye modeling (Świrski & Dodgson, 2013; Dierkes et al., 2018; Popovic et al., 2023; Xiao et al., 2025), For each input frame I, the eyeball is represented as a sphere with parameters

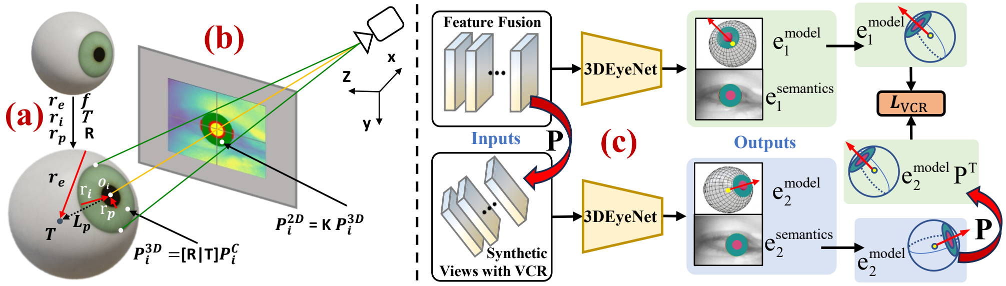

Figure 4: Illustration of structure-aware modeling and view-consistent regularization. (a) The deformed eyeball is reconstructed in the camera coordinate system using the predicted parameters. (b) Projection of the iris and pupil regions onto 2D plane, yielding the corresponding semantic masks. (c) VCR: Enforcing consistency in prediction space and structural edges under simulated rotations.

$E_{\text{para}} = \{r_e, r_i, T, r_p, R\}$, where r_e, r_i, r_p denote eyeball, iris and pupil radii. $T \in \mathbb{R}^3$ indicates the eyeball center in the camera coordinate system, and $R_n \in SO(3)$ describes its orientation. The optical axis is estimated as $g = \frac{o_i - o_e}{\|o_i - o_e\|}$, with o_e and o_i denoting the eyeball and iris center. Camera intrinsics are jointly estimated under a pinhole assumption with $f_x = f_y = f$ and $(c_x, c_y) = (W/2, H/2)$.

Deformation and Projection Following De^{2}Gaze (Xiao et al., 2025), pupil and iris point clouds are generated in the canonical space as concentric disks and rings:

$$\begin{aligned} P_p^C &= \{(r_p\rho\cos(\theta), r_p\rho\sin(\theta), -L_p) \mid \rho \in [0,1], \theta \in [0, 2\pi]\}, \\ P_i^C &= \{(r\cos(\theta), r\sin(\theta),\ -L_p) \mid r = r_p + \rho(r_i - r_p), \rho \in [0,1], \theta \in [0, 2\pi]\}. \end{aligned} \tag{1}$$

These 3D points are transformed to camera space by $[R|T]$ and projected with intrinsic K onto the 2D image: To render the semantic supervision signals, we transform the canonical point clouds into the camera coordinate system by applying the predicted rotation R and translation T: $P_p^{2D} = K[R|T]P_p^C$, $P_i^{2D} = K[R|T]P_i^C$. As shown in Fig. 4(b), only iris/pupil edges are supervised to reduce cost. The MGR branch with structure-aware regression enforces physically grounded predictions, interpretable feedback, and improved robustness to sparse supervision and domain shifts.

3.4 View-Consistent Regularization

Inspired by 3DGazeNet (Ververas et al., 2024) which enforces multi-view consistency with explicit multi-view data, as shown in Fig. 4(c), we introduce View-Consistent Regularization (VCR) that achieves rotation-equivariant consistency without multi-view input. Combining MGR and AGE with adversarial learning, VCR extends consistency beyond geometry to structural and feature levels.

Feature Rotation Let $\mathbf{f}_i \in \mathbb{R}^d$ be the encoder feature. A known rotation $P \in SO(3)$ in the gaze space is mapped to the feature space via a learned linear operator $W \in \mathbb{R}^{d\times 3}$: $\mathbf{f}_i' = \mathbf{f}_i P_f$, $P_f = WPW^\dagger$, where $W^\dagger$ denotes the pseudo-inverse of W, ensuring feature rotation P_f aligns with gaze-space rotation P.
Prediction Consistency. Given $\mathbf{g}_i = \Phi_{dec}(\mathbf{f}_i)$ and $\mathbf{g}_i' = \Phi_{dec}(\mathbf{f}_i')$, where $\Phi(\cdot)$ denotes the decoder. We enforce $\mathbf{g}_i' \approx P\mathbf{g}_i$, so that rotated features yield gaze $\mathbf{g}_i'$ consistent with $\mathbf{g}_i$ after geometric rotation.
Structural Consistency From MGR, the canonical 3D point cloud of the pupil/iris (P_p^{3D}, P_i^{3D}) is rotated by the same rotation P : $\widetilde{P}_p^{3D} = PP_p^{3D}$, $\widetilde{P}_i^{3D} = PP_i^{3D}$. Both point clouds are projected with camera intrinsics K for 2D supervision. $P_p^{2D} = KP_p^{3D}$, $\widetilde{P}_p^{2D} = K\widetilde{P}_p^{3D}$, $P_i^{2D} = KP_i^{3D}$, $\widetilde{P}_i^{2D} = K\widetilde{P}_i^{3D}$.

By jointly constraining rotated features, gaze predictions, and projected eye structures, VCR enforces rotation-equivariance and enhances cross-view generalization.

3.5 Loss Functions

We design a unified loss framework to jointly optimize AGE branch and MGR branch, enhanced by VCR and adversarial branch alignment.

AGE Branch Loss The AGE branch directly regresses gaze from Principle Gaze Features (PGF). The loss is the mean angular error, where the number of samples is N and the loss weight is w_{AGE}:

$$L_{AGE} = w_{AGE}\frac{1}{N}\sum_{n=1}^{N}\arccos\left(\frac{\hat{\mathbf{g}}_{AGE}^n \cdot \mathbf{g}_{GT}^n}{\|\hat{\mathbf{g}}_{AGE}^n\|\|\mathbf{g}_{GT}^n\|}\right). \tag{2}$$

Table 1: Comparison of SG-Gaze with baseline models under cross-dataset and within-dataset evaluation settings. 3D Gaze error is in degrees (lower is better, ▼ denotes reduction, ▲ denotes increase).

Method	$D_{T_1} \to D_S$	$D_{T_1} \to D_L$	$D_{T_1} \to D_{T_2}$	$D_{T_2} \to D_S$	$D_{T_2} \to D_L$	$D_{T_2} \to D_{T_1}$	within $D^*_{T_1}$	within $D^*_{T_2}$
ResNet-18	5.02	6.83	3.20	5.31	6.68	3.29	1.68	1.59
ResNet-18+SG-Gaze	4.10▼18.32%	5.07▼25.77%	2.07▼35.31%	3.93▼25.99%	4.87▼27.10%	2.03▼32.22%	1.11▼33.93%	1.33▼16.35%
ResNet-50	4.04	5.47	2.54	5.16	5.14	3.16	1.32	1.34
ResNet-50+SG-Gaze	3.91▼3.21%	3.75▼31.44%	1.88▼25.78%	4.36▼15.50%	4.21▼18.09%	1.94▼38.61%	1.22▼9.10%	1.09▼18.66%
VGG-16	5.5	7.14	3.61	5.94	7.61	3.94	2.12	2.35
VGG-16+SG-Gaze	5.13▼6.73%	5.2▼27.02%	3.30▼8.59%	5.27▼11.28%	5.33▼29.97%	2.97▼24.62%	2.78▲31.13%	3.13▲33.19%

MGR Branch Loss The MGR branch reconstructs the 3D eyeball and is supervised by both structure projection and gaze. K and N are the total number of edge points and frames.

$$L_{MGR}^{edge} = w_1 \frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} \|P_{nk}^{\text{2D}} - P^{\text{GT,2D}}_{n \arg\min_j \|P_{nk}^{\text{2D}} - P_{nj}^{\text{GT,2D}}\|}\|, \; L_{MGR}^{gaze} = w_2 \frac{1}{N} \sum_{n=1}^{N} \arccos\left(\frac{\hat{\mathbf{g}}^n \cdot \mathbf{g}^n}{\|\hat{\mathbf{g}}^n\| \|\mathbf{g}^n\|}\right). \quad (3)$$

VCR Loss Following Sec 3.4, VCR enforces two view-consistency constraints to ensure robustness under viewpoint changes. $P \in \mathbb{R}^{3\times 3}$ is the ground-truth rotation matrix applied to $\mathbf{g}_i$.

$$L_{VCR}^{gaze} = w'_{\text{gaze}} \tfrac{1}{N} \sum_{i=1}^{N} \|\mathbf{g}'_i - P\mathbf{g}_i\|_2^2, \quad L_{VCR}^{edge} = w'_{\text{proj}} \tfrac{1}{N} \sum_{i=1}^{N} \left(\|P_p^{2D} - \widetilde{P}_p^{2D}\|_2^2 + \|P_i^{2D} - \widetilde{P}_i^{2D}\|_2^2\right). \quad (4)$$

4 Experiments

4.1 Datasets and Preprocessing

Dataset We evaluate on two real-world datasets and one synthetic dataset: **(1) TEyeD**(Fuhl et al., 2021). Comprising 20M+ images from 132 subjects across diverse scenarios, with 2D/3D landmarks, segmentation, gaze vectors, and eye movement labels. We adopt three subject-disjoint splits: D_{T_1} (200k train / 30k test, 16 subjects), D_{T_2} (200k train / 30k test, 16 subjects), and D_S (50k train / 12k test, 39 subjects). These subsets support within-subset and cross-subset performance validation. **(2) LPW** (Tonsen et al., 2016). Collected under daily-life conditions with large illumination, pose, and occlusion variations. D_L: We sample 97k train / 27k test images from 22 subjects, excluding blurred or occluded frames. This dataset serves as a target domain for cross-dataset evaluation under weak 2D edge supervision. **(3) UnityEyes** (Wood et al., 2016). A synthetic dataset with accurate 3D labels and structure annotations. We use 200k images (160k train / 40k test) from 20 virtual subjects, mainly for backbone pretraining, spherical feature validation, and enforcing rotation-consistency, complementing limited real-world labels. All datasets are temporally downsampled from 25 Hz to 6.25 Hz to ensure significant eye movements and avoid redundant frames. **Implementation Details** Input eye images are cropped to 320×240 and normalized. The training dataset is augmented augmented with Gaussian ambiguity (std 1.0–2.0), random noise (0–30%), and horizontal flip (20% probability). Training is performed with batch size 128, each containing 4 consecutive frames. In the MGR branch, structural templates provide weak supervision of iris and pupil boundaries. Specifically, 26 iris edge points are sampled by uniformly distributing angles θ over $\theta \in [0, 0.1\pi] \cup [0.9\pi, 1.1\pi] \cup [1.9\pi, 2\pi]$ with radius $\rho \in [0, 1]$, while 128 pupil contour points are evenly sampled over $[0, 2\pi]$. These serve as pseudo-labels for 2D structural projection. For Isomap, we adopt the Scikit-learn implementation with 300 neighbors in geodesic distance. Optimization uses LAMB (You et al., 2019) with initial learning rate 2×10^{-3} and weight decay 0.02.

4.2 Main Results

Cross-dataset Evaluation We evaluate SG-Gaze by integrating it into different backbones in both domains and within the domain (Table 1). For ResNet-18, SG-Gaze reduces gaze error by up to **35.31%** in $D_{T_1} \to D_{T_2}$ transfer and over **25%** on most tasks. ResNet-50+SG-Gaze achieves **26.77%–38.61%** gains. The slight error increase in VGG within-domain results suggests that SG-Gaze introduces a stronger inductive bias toward structural and cross-view consistency, which may trade off fine-grained fitting in low-capacity models. These results show that SG-Gaze reliably enhances gaze estimation generalization, independent of backbone architecture or domain shift.

Comparison with SOTA Methods We further benchmark SG-Gaze against recent state-of-the-art generalizable approaches, including RAT (Bao et al., 2022), LatentGaze (Lee et al., 2022), FFGaze (Zhang et al., 2017), De^{2}Gaze (Xiao et al., 2025), AGG((Bao & Lu, 2024)) and

Table 2: Cross-domain comparison with state-of-the-art gaze estimation methods. 3D Gaze error is in degrees (lower is better). SG-Gaze achieves the best generalization across unseen domains, benefiting from its structurally and geometrically consistent representation.

Method	$D_{T_1} \to D_S$	$D_{T_1} \to D_{T_2}$	$D_{T_1} \to D_L$	$D_{T_2} \to D_{T_1}$	$D_{T_2} \to D_S$	$D_{T_2} \to D_L$	$D_S \to D_{T_1}$	$D_S \to D_{T_2}$	$D_S \to D_L$	$D_L \to D_{T_1}$	$D_L \to D_{T_2}$	$D_L \to D_S$
RAT (Bao et al., 2022)	5.82	2.83	5.73	3.91	5.31	5.97	6.18	5.29	4.79	5.68	4.59	4.83
Latentgaze (Lee et al., 2022)	5.5	3.14	5.81	3.61	5.94	5.83	5.11	4.99	4.83	5.12	4.35	5.02
FFGaze (Zhang et al., 2017)	5.04	3.47	5.44	3.54	5.16	5.44	4.96	5.16	5.24	5.32	4.34	4.77
PureGaze (Cheng et al., 2022)	4.5	2.94	4.90	2.61	4.94	5.71	4.61	5.94	5.21	5.03	4.35	4.97
AGG (Bao & Lu, 2024)	4.0	2.21	4.33	2.15	4.21	4.94	3.44	4.53	4.21	4.02	3.58	4.44
De^{2}Gaze (Xiao et al., 2025)	4.9	2.88	5.10	3.22	5.02	5.88	4.79	5.25	4.97	5.10	4.72	5.33
Baseline	5.02	3.20	6.83	3.29	5.31	6.68	5.21	5.92	5.41	5.32	4.77	5.63
ResNet18+SG-Gaze	4.10	2.07	5.07	2.03	**3.93**	4.87	3.41	4.36	4.43	4.12	**3.24**	4.23
ResNet50+SG-Gaze	**3.91**	**1.88**	**3.75**	**1.94**	4.36	**4.21**	**3.29**	**4.04**	**4.18**	**3.87**	3.35	**4.01**

Methods	Backbone	Params /(M)	FLOPs /(G)	Loss	TEyeD-D_{T_1} 3D gaze [°]↓	2D gaze [°]↓	Sem. Iou	2D eye cent.[px]↓
(Fuhl et al., 2021)	ResNet50	26.08	24.97	Gaze	1.88	6.90	N/A	N/A
(Kim et al., 2019)	CNN	0.16	0.14	Gaze	3.65	8.34	N/A	N/A
Transformer-based	ResNet18	15.53	14.83	Gaze	1.57	6.36	N/A	N/A
(Vaswani, 2017)	ResNet50	27.66	24.98	Gaze	1.77	6.40	N/A	N/A
QueryDETR	ResNet18	16.36	18.33	Gaze	3.12	8.01	N/A	N/A
(Carion et al., 2020)	ResNet50	29.89	28.42	Gaze	3.08	7.86	N/A	N/A
Nikola et al.	ResNet50	28.56	26.44	Gaze	1.04	7.40	N/A	N/A
	ResNet50	28.56	26.44	Sem.	20.16	39.10	92.5%	11.41
(Popovic et al., 2023)	ResNet50	28.56	26.44	S+G+C	1.21	10.39	91.4%	2.02
De^{2}Gaze	ResNet18	14.48	13.56	Gaze	**0.54**	**5.43**	N/A	N/A
	ResNet18	14.48	13.56	Sem.	20.12	39.09	**94.2%**	11.41
(Xiao et al., 2025)	ResNet18	14.48	13.56	S+G+C	0.96	7.6	93.4%	**1.52**
SG-Gaze (**Ours**)	ResNet18	**11.68**	**11.24**	Gaze	0.94	6.22	N/A	N/A
	ResNet18	11.68	11.24	Sem.	21.12	39.69	93.3%	12.22
	ResNet18	11.68	11.24	S+G+C	1.11	8.82	92.1%	1.88

Table 3: The table reports backbone, Params, FLOPs and loss settings. SG-Gaze achieves the best trade-off between accuracy, efficiency and semantic consistency without domain shift.

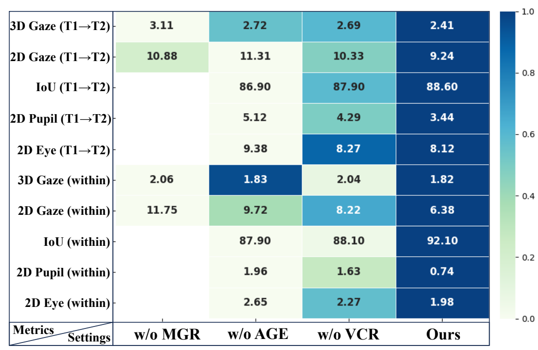

Figure 5: Ablation study heatmap showing the impact of removing different modules (AGE, MGR, VCR) on gaze and semantic metrics, Darker color indicates better performance.

PureGaze (Cheng et al., 2022). As shown in Table 2, SG-Gaze achieves the best or second-best accuracy in nearly all 12 cross-domain transfer tasks. For example, ResNet-50+SG-Gaze reduces the error to **1.88°** on $D_{T_1} \to D_{T_2}$ and **1.94°** on $D_{T_2} \to D_{T_1}$, outperforming the strongest baseline (AGG) by an average of **7.7%**. SG-Gaze also shows clear advantages in challenging scene transfers such as $D_S \to D_{T_1}$ and $D_S \to D_{T_2}$. These results highlight that SG-Gaze not only enhances backbone models but also surpasses existing SOTA methods, demonstrating generalization across diverse domains.

Within-dataset Evaluation We further evaluate SG-Gaze on TEyeD-D_{T_1} without cross-domain transfer, comparing against several traditional methods. As shown in Table 3, SG-Gaze achieves consistently superior performance in multiple evaluation metrics (0.94° for ResNet18 backbone), while keeping the model lightweight (11.68M parameters, 11.24G FLOPs). Moreover, beyond gaze prediction, SG-Gaze also shows improvements in semantic edge IoU and 2D eyeball center localization, indicating that our design enhances both geometric consistency and structural supervision.

4.3 Ablation Studies

We perform ablations on two TEyeD subsets (D_{T_1}, D_{T_2}) to evaluate the contribution of each component by selectively removing or replacing modules. The results are summarized in Table 4.

(1) The Effect of Analytical Gaze Estimation Removing the AGE branch eliminates physically interpretable analytical supervision (*w/o* AGE), leading to a notable error increase to **2.57°** on $D_{T_1} \to D_{T_2}$. Although semantic and center-based constraints still offer auxiliary guidance, the model loses explicit geometric supervision that aligns predicted gaze with analytical eye geometry. This confirms the necessity of integrating analytical geometric priors with data-driven learning. **(2) The Effect of Model-Guided Reconstruction** Removing the MGR branch (*w/o* MGR) leads to the most severe degradation, since no 3D edge projection loss is applied and the structural cues from iris/pupil edges are absent. For instance, in the cross-domain setting $D_{T_1} \to D_{T_2}$, the 3D gaze error increases from **1.88°** to **3.11°**. This confirms that enforcing edge-based structural constraints is crucial for recovering geometric consistency and achieving robust gaze estimation. **(3) The Effect of View-Consistent Regularization** Discarding the VCR module (*w/o* VCR) weakens robustness against cross-view perturbations. The 3D gaze error rises from **1.88°** to **2.07°** in cross-domain evaluation, showing that rotation-consistency regularization not only augments the training distribution with view perturbations but also enhances cross-view generalization. **(4) Ablation Heatmap Visualization** As shown in Fig. 5, the complete model consistently outperforms all ablated versions

Table 4: Ablation study on two TEyeD subsets. We evaluate the complete SG-Gaze model against variants with AGE, MGR, or VCR removed. The full model achieves the best performance across all metrics, demonstrating that each component contributes positively to accuracy and robustness.

Settings	Loss	$D_{T_1} \rightarrow D_{T_2}$					within $D^*_{T_1}$				
		3D gaze [°]↓	2D gaze [°]↓	Sem. Iou	2D pupil cent.[px]↓	2D eye cent.[px]↓	3D gaze [°]↓	2D gaze [°]↓	Sem. Iou	2D pupil cent.[px]↓	2D eye cent.[px]↓
w/o MGR Module	Gaze	3.11	10.88	N/A	N/A	N/A	2.86	11.75	N/A	N/A	N/A
w/o AGE Module	Gaze	2.57	9.76	N/A	N/A	N/A	1.83	10.32	N/A	N/A	N/A
	Sem. + Gaze + Cent.	2.72	11.31	86.9%	5.12	9.38	2.54	9.72	87.9%	1.96	2.65
w/o VCR	Gaze	2.07	9.36	N/A	N/A	N/A	1.53	8.12	N/A	N/A	N/A
	Sem. + Gaze + Cent.	2.69	10.33	87.9%	4.29	8.27	2.04	9.22	88.1%	1.63	2.27
SG-Gaze (**Ours**)	Gaze	**1.88**	**8.02**	N/A	N/A	N/A	**0.94**	**6.22**	N/A	N/A	N/A
	Sem. + Gaze + Cent.	2.41	9.24	**88.6%**	**3.44**	**8.12**	1.11	8.82	**92.1%**	**0.74**	**1.98**

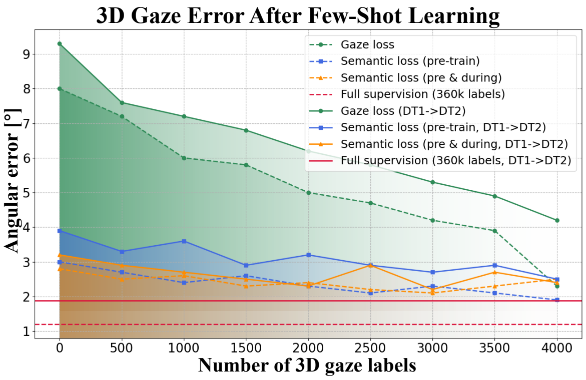

Figure 6: Few-shot fine-tuning results showing that semantic pre-training enables consistent performance gains within and across datasets, even with very limited gaze annotations.

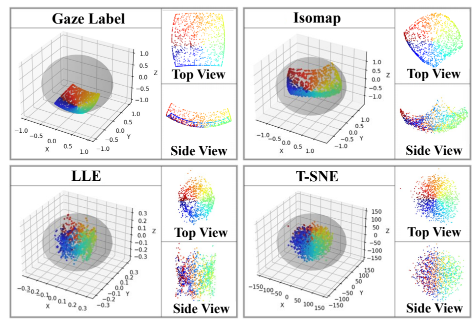

Figure 7: 3D visualization of gaze representations on UnityEyes after dimensionality reduction (Isomap, t-SNE, and LLE). only Isomap preserves the global manifold and aligns well with GT gaze directions.

across metrics. These results highlight that multi-task joint training is essential for balanced performance. Furthermore, semantic IoU and 2D localization achieve the best results, confirming that our unified framework improves both geometric fidelity and physical consistency.

4.4 Efficient Learning under Limited Supervision

(1) Few-Shot Adaptation Accurate gaze estimation is limited by the scarcity of reliable 3D gaze annotations in head-mounted datasets (Kim et al., 2019; Fuhl et al., 2021), while iris and pupil masks are abundant and easier to label. To leverage this imbalance, we adopt a two-stage scheme: pre-train on large-scale semantic labels, then fine-tune with few 3D gaze samples. As shown in Fig. 6, training from scratch with limited annotations performs poorly, whereas semantic pre-training provides a reasonable zero-shot baseline. Finally, even with only 1% labeled gaze samples per subject, combining pre-training with light fine-tuning yields the largest improvement. **(2) Choice of Distance Metrics** We evaluate three dimensionality reduction methods to project features f_i into 3D space. As shown in Fig. 7, gaze labels form a smooth spherical manifold. Isomap (Tenenbaum et al., 2000) best preserves global geometry and orientation. LLE (Roweis & Saul, 2000) retains local neighborhoods but collapses global structure. T-SNE (Maaten & Hinton, 2008) distorts geometry and shows sensitivity to noise. Hence, Isomap is chosen for learning gaze representation.

4.5 Qualitative Results

(1) Effect of VCR on Gaze Distribution We visualize pitch–yaw distributions with and without VCR. As shown in Fig. 8(a), models without VCR yield clustered predictions biased toward dominant training views, limiting gaze coverage. With VCR, the distribution becomes more diverse and uniform, especially at extreme angles, reducing domain gaps between controlled and in-the-wild datasets, and improving generalization by enforcing rotational equivariance (Fig. 8(b)). **(2) Prediction and Rendering Visualization** The 3D eye model and estimated gaze direction are projected onto the 2D image plane. As shown in Fig. 9, the predicted 3D gaze vectors (Fig. 9(b)) remain consistent with the eye's viewing direction across frames, confirming estimation reliability. The

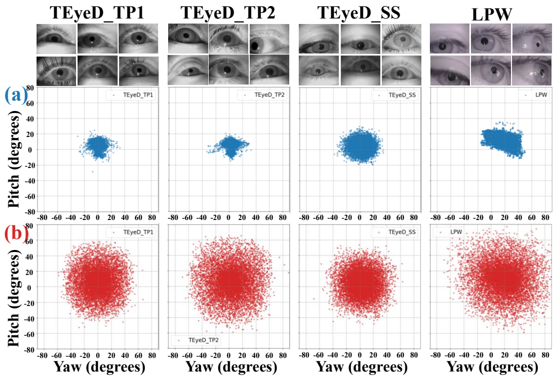

Figure 8: Distribution of gaze. (a) Without VCR, predictions concentrate within a narrow range. (b) With VCR, distributions become more diverse and evenly spread, covering a larger field-of-view.

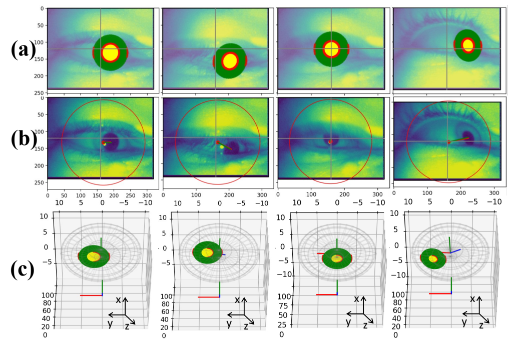

Figure 9: (a) 2D projection of pupil/iris from MGR branch (red areas represent the edge sampling points). (b) Predicted gaze (green) vs. ground truth (red). (c) Reconstructed 3D eyeball model in camera space.

reconstructed 3D eye model (Fig. 9(c)) is well aligned with pupil and iris projections (Fig. 9(a)), demonstrating accurate structural recovery and geometrically plausible 2D rendering.

5 Related Work

Model-based Methods Model-based approaches reconstruct eyeball's anatomical structure (Chen et al., 2008; Hennessey et al., 2006; Wood & Bulling, 2014) using infrared glints, corneal reflections (Hansen & Ji, 2009), or parametric 3D models (Smith et al., 2020; Popovic et al., 2023). Differentiable rendering (Wood et al., 2019; Yoon et al., 2021), neural implicit surfaces (Guo et al., 2022), and hybrid pipelines (Zhang et al., 2021) have improved geometric fidelity and robustness. Recent works extend model-based estimation with representation learning and adaptability: De^{2}Gaze (Xiao et al., 2025) proposed a 3D eye tracking method using deformable and decoupled representations with lightweight time-series. Few-shot personalization (Wang et al., 2022a) and uncertainty-aware fitting (Kim et al., 2021) further enhance cross-subject adaptability.

Appearance-based Methods Deep learning has driven most recent progress in gaze estimation, leveraging large-scale datasets (Krafka et al., 2016; Zhang et al., 2020; 2015; Kellnhofer et al., 2019; Zhang et al., 2017). Models predictions from eye crops (Zhang et al., 2015; Cheng et al., 2020b), full-face images (Bao et al., 2021; Balim et al., 2023; Chen & Shi, 2018; Cheng et al., 2020a; Krafka et al., 2016; Zhang et al., 2017), or their fusion (Krafka et al., 2016; Park et al., 2020). Leveraging two-eye asymmetry has been shown to improve prediction accuracy (Chen & Shi, 2018; Cheng et al., 2020b). Recent approaches employ attention mechanisms, hierarchical and multi-stream architectures, as well as transformer-based backbones (Qin et al., 2025; Yin et al., 2024; Miyato et al., 2023). **Cross-domain Gaze Estimation** Existing methods include unsupervised domain adaptation (Lahiri et al., 2018; Zhang et al., 2022a), contrastive objectives (Wang et al., 2022b) and collaborative learning with outlier guidance (Liu et al., 2021). But these methods require target-domain data. Techniques include rotation-consistent regularization (Bao et al., 2022), source-domain feature purification (Cheng et al., 2022), style perturbation (Zhou et al., 2024), and synthetic imagery (Zhang et al., 2022b). Recent work incorporate geometric constraints and 3D structure. 3D eye mesh regression (Ververas et al., 2024) and geometry projections with spherical training (Bao & Lu, 2024) enhance cross-domain robustness without target-domain access.

6 Conclusion

In this work, we proposed SG-Gaze, a physically guided dual-branch framework that integrates analytical gaze estimation, model-guided reconstruction, and view-consistent regularization. Our approach enforces structural and geometric consistency while promoting rotation-equivariant feature learning, yielding accurate and generalizable 3D gaze representations. Extensive experiments demonstrate its strong cross-domain generalization and improved gaze diversity, highlighting the benefits of incorporating physical priors into learned representations. SG-Gaze provides a foundation for scalable, few-shot, and personalized gaze estimation. Future work will explore learning temporal and personalized gaze representations from larger in-the-wild datasets, aiming to improve cross-subject generalization and adaptation in real-world AR/VR environments.

Ethics Statement

All authors have read and agree to abide by the ICLR Code of Ethics[1]. This work focuses on 3D gaze estimation and reconstruction using eye images from publicly available datasets. No experiments involve human or animal subjects directly, and no personally identifiable or private information is collected. All datasets used (e.g., TEyeD (Fuhl et al., 2021), LPW (Tonsen et al., 2016), UnityEyes (Wood et al., 2016)) are widely adopted in the research community and comply with their respective licenses. To further ensure ethical compliance, we do not release any raw personal eye-tracking data, but only model code and trained weights for reproducibility.

The potential benefits of this research include applications in AR/VR interaction, assistive technologies for people with motor impairments, and clinical support tools (e.g., vision diagnostics). However, we acknowledge that gaze tracking technologies can raise concerns regarding privacy, surveillance, and potential misuse in sensitive contexts. Our method improves accuracy and generalizability, but it should be applied responsibly and only in scenarios that respect user consent and privacy.

We encourage the research community to use our contributions solely for beneficial applications and in accordance with ethical guidelines. We commit to releasing our code, model architectures, and training details in line with the standards of reproducibility and transparency. No conflicts of interest or external sponsorships that might have influenced this work exist.

Reproducibility Statement

We take reproducibility seriously and have taken multiple measures to ensure that all aspects of our work can be replicated by other researchers. First, we provide anonymized source code as supplementary material, which includes the complete implementation of the proposed SG-Gaze framework, detailed configuration files, and training/evaluation scripts.

Second, we explicitly report all hyperparameters such as optimizer type, learning rate, weight decay, batch size, training epochs, and initialization strategies, making it possible to reproduce our training schedules.

Third, for our proposed View-Consistent Regularization (VCR), we include the complete mathematical formulation, proof sketches, and implementation details in the Appendix to clarify all theoretical claims.

Fourth, additional ablation studies, sensitivity analyses, and visualizations are provided in the supplementary material to support the design choices and robustness of the method.

Finally, we encourage further validation by including scripts to reproduce all figures and tables reported in the paper.

Together, these efforts ensure that our results are reproducible and provide a solid foundation for future research on interpretable and generalizable gaze estimation.

Use of Large Language Models (LLMs)

In line with the ICLR 2025 policy on the use of large language models (LLMs), we provide a transparent account of how such tools were utilized in the preparation of this paper. We stress that the scientific contributions, including the formulation of the problem, the design of the method, the theoretical derivations, and the experimental implementation, were entirely conceived and executed by the authors. The role of LLMs was restricted to auxiliary support as outlined below:

- **Writing and Editing:** LLMs (e.g., GPT-based models) were used to assist with grammar checking, sentence restructuring, and improving the clarity and readability of the manuscript. This included rephrasing certain technical descriptions for conciseness and consistency of style across sections.

[1] `https://iclr.cc/public/CodeOfEthics`

- **Literature Organization:** LLMs were used to generate initial summaries of related work from bibliographic entries provided by the authors. These summaries were manually reviewed, corrected, and integrated into the final Related Work section by the authors.
- **Proofreading and Consistency Checking:** LLMs assisted in detecting potential inconsistencies in notation and formatting across the paper (e.g., consistent use of symbols for loss functions, dataset names, and evaluation metrics).

Importantly, LLMs were *not* used to generate research ideas, design experiments, propose model architectures, perform data analysis, or interpret results. All conceptual insights, technical innovations, and conclusions are solely attributed to the authors. We include this statement to ensure transparency and compliance with the ICLR Code of Ethics regarding responsible use of LLMs.

References

Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.

Haldun Balim, Seonwook Park, Xi Wang, Xucong Zhang, and Otmar Hilliges. Efe: End-to-end frame-to-gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2688–2697, 2023.

Yiwei Bao and Feng Lu. From feature to gaze: A generalizable replacement of linear layer for gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1409–1418, 2024.

Yiwei Bao, Yihua Cheng, Yunfei Liu, and Feng Lu. Adaptive feature fusion network for gaze tracking in mobile tablets. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 9936–9943. IEEE, 2021.

Yiwei Bao, Yunfei Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with rotation consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4207–4216, 2022.

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 *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–13, 2020.

Paola Natalia Cañas, Alexander Diez, David Galvañ, Marcos Nieto, and Igor Rodríguez. Occlusion-aware driver monitoring system using the driver monitoring dataset. *arXiv preprint arXiv:2504.20677*, 2025.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Jixu Chen, Yan Tong, Wayne Gray, and Qiang Ji. A robust 3d eye gaze tracking system using noise reduction. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, pp. 189–196, 2008.

Zhaokang Chen and Bertram E Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pp. 309–324. Springer, 2018.

Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 10623–10630, 2020a.

Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272, 2020b.

Yihua Cheng, Yiwei Bao, and Feng Lu. Puregaze: Purifying gaze feature for generalizable gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 436–443, 2022.

Yihua Cheng, Yaning Zhu, Zongji Wang, Hongquan Hao, Yongwei Liu, Shiqing Cheng, Xi Wang, and Hyung Jin Chang. What do you see in vehicle? comprehensive vision solution for in-vehicle gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1556–1565, 2024.

Kai Dierkes, Moritz Kassner, and Andreas Bulling. A novel approach to single camera, glint-free 3d eye model fitting including corneal refraction. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pp. 1–9, 2018.

Wolfgang Fuhl, Gjergji Kasneci, and Enkelejda Kasneci. 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. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 367–375. IEEE, 2021.

Xiaoyu Guo, Yang Cheng, and Xiang Zhou. Nerfeye: Neural radiance fields for eye gaze and shape estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 215–231. Springer, 2022.

Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Craig Hennessey, Borna Noureddin, and Peter Lawrence. A single camera eye-gaze tracking system with free head motion. In *Proceedings of the 2006 symposium on Eye tracking research & applications*, pp. 87–94, 2006.

Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6912–6921, 2019.

Joohwan Kim, Michael Stengel, Alexander Majercik, Shalini De Mello, David Dunn, Samuli Laine, Morgan McGuire, and David Luebke. 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.

Minsoo Kim, Jihoon Park, and Tae-Kyun Kim. Uncertainty-aware 3d gaze estimation with adaptive fitting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11210–11219, 2021.

Robert Konrad, Anastasios Angelopoulos, and Gordon Wetzstein. Gaze-contingent ocular parallax rendering for virtual reality. *ACM Transactions on Graphics (TOG)*, 39(2):1–12, 2020.

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*, pp. 2176–2184, 2016.

Avisek Lahiri, Abhinav Agarwalla, and Prabir Kumar Biswas. Unsupervised domain adaptation for learning eye gaze from a million synthetic images: An adversarial approach. In *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 1–9, 2018.

Isack Lee, Jun-Seok Yun, Hee Hyeon Kim, Youngju Na, and Seok Bong Yoo. Latentgaze: Cross-domain gaze estimation through gaze-aware analytic latent code manipulation. In *Proceedings of the asian conference on computer vision*, pp. 3379–3395, 2022.

Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3835–3844, 2021.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Takeru Miyato, Bernhard Jaeger, Max Welling, and Andreas Geiger. Gta: A geometry-aware attention mechanism for multi-view transformers. *arXiv preprint arXiv:2310.10375*, 2023.

Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges. Towards end-to-end video-based eye-tracking. In *European conference on computer vision*, pp. 747–763. Springer, 2020.

Nikola Popovic, Dimitrios Christodoulou, Danda Pani Paudel, Xi Wang, and Luc Van Gool. Model-aware 3d eye gaze from weak and few-shot supervisions. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 746–751. IEEE, 2023.

Jiawei Qin, Xucong Zhang, and Yusuke Sugano. Unigaze: Towards universal gaze estimation via large-scale pre-training. *arXiv preprint arXiv:2502.02307*, 2025.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Bradley Smith, Michael Stengel, and Thomas Greer. A model-based gaze estimation method using monocular and stereo eye images. *Computers & Graphics*, 89:45–55, 2020.

Lech Świrski and Neil A Dodgson. A fully-automatic, temporal approach to single camera, glint-free 3d eye model fitting. *Proceedings of ECEM*, 2013, 2013.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Yunus Terzioğlu, Bilge Mutlu, and Erol Şahin. Designing social cues for collaborative robots: the role of gaze and breathing in human-robot collaboration. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pp. 343–357, 2020.

Marc Tonsen, Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Labelled pupils in the wild: a dataset for studying pupil detection in unconstrained environments. In *Proceedings of the ninth biennial ACM symposium on eye tracking research & applications*, pp. 139–142, 2016.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Evangelos Ververas, Polydefkis Gkagkos, Jiankang Deng, Michail Christos Doukas, Jia Guo, and Stefanos Zafeiriou. 3dgazenet: Generalizing 3d gaze estimation with weak-supervision from synthetic views. In *European Conference on Computer Vision*, pp. 387–404. Springer, 2024.

Chen Wang, Yifan Xu, and Feng Lu. Few-shot gaze estimation via personalized calibration networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Yaoming Wang, Yangzhou Jiang, Jin Li, Bingbing Ni, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Teng Li. Contrastive regression for domain adaptation on gaze estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19376–19385, 2022b.

Erroll Wood and Andreas Bulling. Eyetab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the symposium on eye tracking research and applications*, pp. 207–210, 2014.

Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 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.

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, and Jamie Shotton. Rendering eye images for gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4667–4676, 2019.

Yunfeng Xiao, Xiaowei Bai, Baojun Chen, Hao Su, Hao He, Liang Xie, and Erwei Yin. Deˆ 2gaze: Deformable and decoupled representation learning for 3d gaze estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3091–3100, 2025.

Pengwei Yin, Jingjing Wang, Guanzhong Zeng, Di Xie, and Jiang Zhu. Lg-gaze: Learning geometry-aware continuous prompts for language-guided gaze estimation. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Yoonsik Yoon, Seonghyeon Park, and Tae-Kyun Kim. Differentiable gaze estimation with 3d eye mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10035–10045, 2021.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

Hanlin Zhang, Xinming Wang, Weihong Ren, Ruihan Lin, and Honghai Liu. Outlier constrained unsupervised domain adaptation algorithm for gaze estimation. In *International Conference on Intelligent Robotics and Applications*, pp. 349–359. Springer, 2022a.

Jiawei Zhang, Yifan Liu, and Andreas Bulling. Hybrid eye model reconstruction using sparse landmarks and dense priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13590–13599, 2021.

Mingfang Zhang, Yunfei Liu, and Feng Lu. Gazeonce: Real-time multi-person gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4197–4206, 2022b.

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*, pp. 4511–4520, 2015.

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 51–60, 2017.

Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Ethxgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European conference on computer vision*, pp. 365–381. Springer, 2020.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Mixstyle neural networks for domain generalization and adaptation. *International Journal of Computer Vision*, 132(3):822–836, 2024.

Appendix

Section A introduces our demo video information. A live demo of our method can be found at this link. Section B describes the implementation process of the method in more details. We have carried out additional experiments, and the results will be explained in Section C. Finally, in Section D, we will discuss the limitations of the method and future work. Our code will be released on GitHub upon acceptance.

A Demo Video

To better illustrate our proposed method, we provide a demo video as supplementary material. The video demonstrates the complete pipeline of our 3D gaze estimation framework, including raw eye image input, 2D rendering, 3D eyeball reconstruction, and mesh visualization. Specifically:

- **Original Image:** The raw infrared eye image captured by the tracker.
- **2D Rendering:** Projection of the reconstructed eyeball and pupil region onto the image plane.
- **Eye Tracker + Sphere:** Visualization of the eyeball sphere model aligned with the tracker coordinate system.

- **3D Eyeball:** Fusion of image appearance with the 3D eyeball model for interpretable gaze analysis.
- **3D Mesh:** Geometric mesh reconstruction of the eyeball with anatomical and structural priors.

The video highlights how our model integrates both appearance features and structural constraints, ensuring interpretable and physically consistent gaze estimation.

B Methodological Details

B.1 Render Semantics (MGR branch)

We begin by generating the template point clouds for both the pupil and the iris using polar coordinates. The process involves discretizing the angles and radii, and then converting the polar coordinates into 3D Cartesian coordinates.

Generate Angle Grid We generate a set of angles in the range $[0, 2\pi)$ for each batch and frame. This is done by discretizing the angle space into N_{angles} points:

$$\theta_i = \frac{2\pi i}{N_{\text{angles}}}, \quad \text{for} \quad i = 0, 1, 2, \ldots, N_{\text{angles}} - 1, \tag{5}$$

this results in an angle grid of shape $[B, N_{\text{angles}}]$, where B is the batch size multiplied by the number of frames.

Generate Radius Grids The radius of the pupil and iris are discretized into N_{radius} points. For the pupil, the radii range from 0 to r_{pupil}, while for the iris, they range from r_{pupil} to r_{iris}. The radii are generated as:

$$\begin{aligned} r^i_{\text{pupil}} &= \frac{i}{N_{\text{radius}}} r_{\text{pupil}}, \\ r^i_{\text{iris}} &= \frac{i}{N_{\text{radius}}} (r_{\text{iris}} - r_{\text{pupil}}) + r_{\text{pupil}}, \\ &\text{for} \quad i = 0, 1, 2, \ldots, N_{\text{radius}} - 1, \end{aligned} \tag{6}$$

this generates two radius grids, one for the pupil and one for the iris, both of shape $[B, N_{\text{radius}}]$.

Convert Polar Coordinates to Cartesian Coordinates. We convert the polar coordinates into 3D Cartesian coordinates for both the pupil and the iris. For each pair of radius r and angle θ, the Cartesian coordinates (x, y, z) are computed as:

$$x = r \cdot \cos(\theta), \quad y = r \cdot \sin(\theta). \tag{7}$$

For the pupil and iris point clouds, the z-coordinate is set to a fixed value, determined by the distance from the camera, L_p, and inverted to place the point clouds in front of the camera:

$$z_{\text{pupil}} = z_{\text{iris}} = -L_p \tag{8}$$

Thus, the final 3D coordinates for the pupil and iris are:

$$\begin{aligned} P_{\text{pupil}} &= \left(x_{\text{pupil}}, y_{\text{pupil}}, z_{\text{pupil}}\right), \\ P_{\text{iris}} &= \left(x_{\text{iris}}, y_{\text{iris}}, z_{\text{iris}}\right). \end{aligned} \tag{9}$$

The resulting point clouds have shapes $[B, N_{\text{angles}} \times N_{\text{radius}}, 3]$.

B.2 AGE branch details

A key objective of AGE branch is to train the feature extractor F_{θ_1} under the guidance of spherical fitting, so that gaze features are aligned with the physical definition of eye rotations. However, directly integrating the Isomap (Tenenbaum et al., 2000) algorithm into backpropagation is computationally prohibitive: the time complexity of Isomap is $O(N^2 \log N)$ and the memory complexity is

$O(N^2)$, where N denotes the number of samples. Processing hundreds of thousands of gaze features with Isomap is thus infeasible in both time and memory.

Isometric Propagator (IP) To overcome this limitation, we introduce the Isometric Propagator (IP) following previous studies (Bao & Lu, 2024), a lightweight three-layer MLP $IP_{\theta_3}(\cdot)$ that parameterizes the Isomap algorithm. The IP is trained to approximate Isomap embeddings during an initialization phase. Specifically, we freeze the parameters of the pretrained CNN F_{θ_1} and train IP_{θ_3} to regress the Isomap outputs from its input features:

$$\min_{\theta_3} \frac{1}{N} \sum_{i=1}^{N} L_1 \left(\mathrm{Isomap}(f_i),\, IP_{\theta_3}(f_i) \right), \tag{10}$$

where $f_i = F_{\theta_1}(x_i)$ are CNN features and $\mathscr{L}_1$ denotes the L1 loss.

Retraining the Feature Extractor After training, we freeze the parameters of IP_{θ_3} and replace Isomap with it for training the feature extractor. Sphere-Fitting Training objective is then defined as:

$$\min_{\theta_1} \frac{1}{N} \sum_{i=1}^{N} L_1 \left(\hat{e}_i,\, IP_{\theta_3}(F_{\theta_1}(x_i)) \right), \tag{11}$$

where $\hat{e}_i$ denotes the ground-truth embedding derived from spherical fitting. The Isometric Propagator is only used during source-domain training.

Inference At test time, gaze is estimated through Isomap and Spherical Fitting:

$$g_i = SF_{\theta_s}\left(\mathrm{Isomap}(F_{\theta_1}(x_i))\right). \tag{12}$$

This training strategy directly optimizes gaze features according to their geometric relation with spherical fitting, ensuring physical interpretability while maintaining computational feasibility.

B.3 Training Details

We use ResNet-18/50 as backbones for fair comparison, followed by our dual-branch decoder. SG-Gaze is trained on NVIDIA A100 GPU, with a batch size of 128. We set the initial weights of projection edge loss λ_{edge}, eyeball center loss λ_{eye}^{center} and pupil center loss λ_{pupil}^{center} to 0.15. The initial weight of that two 3D gaze loss λ_{gaze}^{L2} and $\lambda_{gaze}^{cos-sin}$ are set to 2.5. The training process is terminated at 160 epochs. The Sphere-Fitting Training is 20 epochs, while the IP is trained for 100 epochs on 10000 randomly selected samples.

B.4 View-Consistent Regularization (VCR)

Rotation parameterization We use the right-handed camera coordinate system and compose the 3D viewpoint rotation as $P(\alpha,\beta,\gamma) = R_z(\gamma)R_y(\beta)R_x(\alpha) \in \mathrm{SO}(3)$, where α,β,γ denote pitch (x-axis), yaw (y-axis) and roll (z-axis), respectively. Unless otherwise stated, angles are in degrees.

Mixture-of-ranges sampling To simultaneously cover near-view perturbations and far-view shifts while keeping training stable, we sample (α,β,γ) from a two-component mixture:

$$(\alpha,\beta,\gamma) \sim (1-\lambda)\,\Omega_{\mathrm{loc}} \,+\, \lambda\,\Omega_{\mathrm{glob}},$$

with $\lambda = 0.2$. The local component focuses on small perturbations

$$\Omega_{\mathrm{loc}} = U([-12^\circ, 12^\circ]) \times U([-12^\circ, 12^\circ]) \times U([-6^\circ, 6^\circ]), \tag{13}$$

and the global component sweeps a wider FoV, dataset-aware:

$$\begin{aligned} \Omega_{\mathrm{glob}} =& U([-A_{\max}, A_{\max}]) \times \\ & U([-B_{\max}, B_{\max}]) \times U([-G_{\max}, G_{\max}]). \end{aligned} \tag{14}$$

For UnityEyes pretraining we use $(A_{\max}, B_{\max}, G_{\max}) = (60^\circ, 50^\circ, 15^\circ)$; for TEyeD/LPW fine-tuning we use $(35^\circ, 30^\circ, 10^\circ)$. This setting matches the broader synthetic coverage while avoiding excessive roll in head-mounted real captures. We further adopt a two-stage curriculum: for the first E=10 epochs, λ=0 (local-only), then switch to the above mixture.

AGE branch under rotation (feature-space). Let $f_i \in \mathbb{R}^d$ be the shared-encoder feature of frame I_i. We map the 3D rotation into feature space via the learned linear operator $W \in \mathbb{R}^{d\times 3}$ and its pseudo-inverse $W^\dagger$:

$$f_i' = f_i P_f, \qquad P_f = W P W^\dagger.$$

The analytical decoder Φ_{dec} (shared for original/rotated features) predicts

$$g_i = \Phi_{\text{dec}}(f_i), \qquad g_i' = \Phi_{\text{dec}}(f_i'),$$

and VCR enforces rotation-equivariant gaze prediction

$$L_{\text{VCR}}^{\text{gaze}} = w_{\text{gaze}}' \frac{1}{N} \sum_{i=1}^{N} \| g_i' - P g_i \|_2^2.$$

We do *not* re-encode images after rotation; gradients flow through W, $W^\dagger$ and Φ_{dec}.

MGR branch under rotation (structure-space). From the MGR branch we have canonical pupil/iris point clouds PC_p, PC_i and their camera-space instances P_p^{3D}, P_i^{3D} obtained by the predicted pose $[R|T]$ and then projected by K to P_p^{2D}, P_i^{2D} (see main text). VCR applies the *same* geometric rotation P to the camera-space point clouds (no re-run of MGR):

$$\widetilde{P}_p^{3D} = P P_p^{3D}, \quad \widetilde{P}_i^{3D} = P P_i^{3D}, \quad \widetilde{P}_p^{2D} = K \widetilde{P}_p^{3D}, \quad \widetilde{P}_i^{2D} = K \widetilde{P}_i^{3D}.$$

We enforce 2D semantic edge consistency via nearest-neighbor matching:

$$L_{\text{VCR}}^{\text{edge}} = w_{\text{proj}}' \frac{1}{N} \sum_{i=1}^{N} \left(\| P_p^{2D} - \widetilde{P}_p^{2D} \|_2^2 + \| P_i^{2D} - \widetilde{P}_i^{2D} \|_2^2 \right).$$

The total VCR regularization is

$$L_{\text{VCR}} = L_{\text{VCR}}^{\text{gaze}} + L_{\text{VCR}}^{\text{edge}},$$

which used alongside AGE/MGR losses in the joint objective. In practice we share the decoder across original/rotated features (AGE) and reuse the predicted structure for geometric rotation (MGR), which avoids extra backbone passes while enforcing consistent physics across views.

C Additional Experiments

C.1 Effect of Rotation Angles in VCR

Motivation Our View-Consistent Regularization (VCR) applies synthetic viewpoint perturbations during training to enforce rotation-equivariant consistency. The choice of rotation ranges may affect the trade-off between local robustness and global generalization. Here we study the influence of different angle ranges on gaze estimation accuracy.

Experimental setup We compared three rotation ranges for yaw, pitch, and roll axes:

- *Small rotation:* yaw $\in [-10°, 10°]$, pitch $\in [-10°, 10°]$, roll $\in [-5°, 5°]$.
- *Medium rotation:* yaw $\in [-20°, 20°]$, pitch $\in [-15°, 15°]$, roll $\in [-10°, 10°]$.
- *Large rotation:* yaw $\in [-40°, 40°]$, pitch $\in [-30°, 30°]$, roll $\in [-15°, 15°]$.

All other settings followed the main training configuration. We report angular gaze error (degrees) on TEyeD (Fuhl et al., 2021) and LPW (Tonsen et al., 2016) benchmarks.

Discussion. The results in Table 5 show that overly small perturbations fail to simulate diverse viewpoints, while excessively large rotations introduce unrealistic samples. Medium-range perturbations strike the best balance, improving cross-view robustness and cross-domain generalization.

C.2 Effect of the kappa angle between the optical and visual axes.

The normalized optical axis g is defined as the vector from the eyeball center o_e to the iris center o_i, $g = \frac{o_i - o_e}{\|o_i - o_e\|}$. We consider g the approximated gaze vector. Note that we did not model the

Table 5: Effect of different rotation ranges in VCR on gaze estimation accuracy (angular error, degrees). Medium-range perturbations achieve the best trade-off.

Rotation range	$D_{T_1} \to D_{T_2} \downarrow$	$D_{T_1} \to D_S \downarrow$
Small (±10° / ±10° / ±5°)	6.01	3.22
Medium (±20° / ±15° / ±10°)	**3.91**	**1.88**
Large (±40° / ±30° / ±15°)	4.65	2.79

Table 6: Quantitative results of individual training and testing of five subjects on TEyeD dataset. After eliminating the influence of kappa angle, the results show that applying more constraints is beneficial to improve the reconstruction accuracy of 3D eyeball model.

Subject	Loss	TEyeD-subset_A				
		3D gaze [°]↓	2D gaze [°]↓	Sem. Iou	2D pupil cent.[px]↓	2D eye cent.[px]↓
subject1	Gaze	1.21	7.59	N/A	N/A	N/A
	Sem. + Gaze + Cent.	0.85	4.66	86.5%	3.11	2.33
subject2	Gaze	1.32	8.50	N/A	N/A	N/A
	Sem. + Gaze + Cent.	0.99	5.21	87.3%	3.45	1.87
subject3	Gaze	1.50	7.05	N/A	N/A	N/A
	Sem. + Gaze + Cent.	1.07	6.33	88.1%	1.22	1.52
subject4	Gaze	1.47	7.32	N/A	N/A	N/A
	Sem. + Gaze + Cent.	1.11	5.08	86.5%	2.51	2.78
subject5	Gaze	1.34	7.74	N/A	N/A	N/A
	Sem. + Gaze + Cent.	1.01	4.57	86.4%	2.71	2.40

kappa angle offset between the optical and visual axes. In our previous experiments, we put more supervision on the whole eyeball, such as the center of the eyeball and pupil, as well as the edge of the projection. However, the accuracy has declined. To better understand its impact, we conducted controlled experiments. We trained and tested five subjects separately to eliminate the influence of kappa angle difference among different subjects. The experimental results in Tab 6 highlight that subject-dependent kappa offsets can negatively affect gaze estimation accuracy if ignored. Adding more subject-specific supervision constraints improves consistency of eyeball fitting and yields more reliable 3D gaze estimation.

C.3 Sensitivity of MGR to 2D Edge Sparsity

Motivation. The Model-Guided Reconstruction (MGR) branch is supervised via sparse 2D edge points sampled on the pupil and iris contours. In practice, acquiring dense semantic labels can be expensive; therefore we analyze how the number of sampled 2D edge points K affects reconstruction fidelity and gaze estimation performance. This experiment quantifies the trade-off between annotation cost (semantic sparsity) and final accuracy.

Experimental setup.

- **Backbone & training:** We use the same backbone and training hyperparameters as in the main paper (ResNet-18 / ResNet-50 variants as applicable). The training schedule, optimizer, weight decay and loss weights are kept identical to the main experiments to isolate the effect of K.
- **Point sampling:** For a given K we uniformly sample K_p contour points on the pupil and K_i points on the iris such that $K = K_p + K_i$. By default we keep the pupil : iris ratio the same as in the main paper (e.g., $K_p : K_i = 4 : 1$ if the paper uses 128 pupil vs 32 iris).

- **K values tested:** $K \in \{8, 16, 32, 64, 128, 256\}$.
- **Datasets & evaluation:** Experiments are run on TEyeD-D_{T_1} and evaluated on the same test splits used in the main paper. Metrics reported are 3D gaze angular error (degrees),2D edge reprojection error (mean pixel distance), and semantic IoU.

Losses and weights. We keep the same overall objective as in the main paper:

$$\mathscr{L} = \lambda_g \mathscr{L}_{\text{gaze}} + \lambda_e \mathscr{L}_{\text{edge}} + \lambda_v \mathscr{L}_{\text{vcr}}.$$

When varying K we do *not* change λ_e; this isolates the effect of the amount of 2D structural supervision. Reported runs keep $\lambda_g, \lambda_e, \lambda_v$ identical to the main experiments.

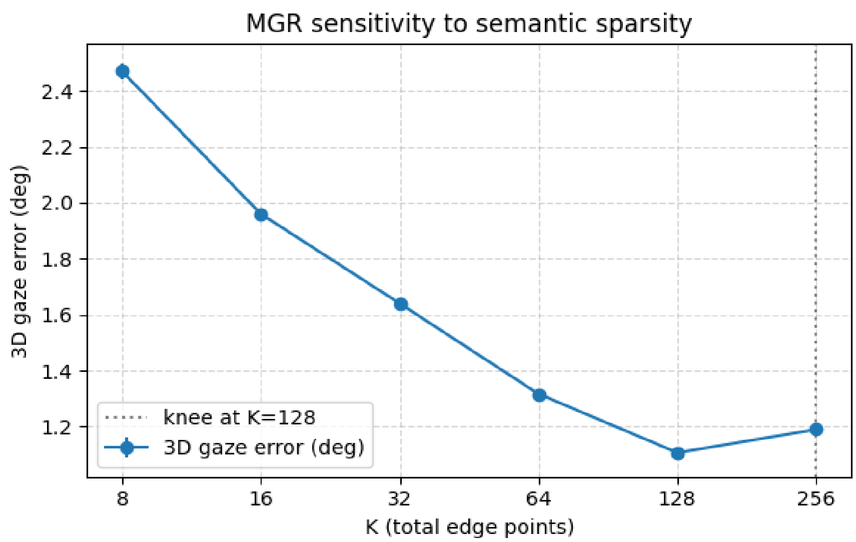

Figure 10: Sensitivity of the MGR branch to the number of 2D edge points K. Increasing K rapidly reduces 3D gaze error up to $K = 128$, beyond which the performance saturates, indicating diminishing returns.

Table 7: Sensitivity of MGR to number of 2D edge points K. For each K we report 3D gaze angular error (deg, lower better), 2D gaze angular error (deg, lower better), and semantic IoU (%, higher better).

K (total points)	3D gaze (°) ↓	2D gaze (°) ↓	Sem. IoU (%) ↑
8	2.47 ± 0.08	15.62 ± 0.25	78.5 ± 0.6
16	1.96 ± 0.07	13.24 ± 0.21	83.7 ± 0.5
32	1.64 ± 0.05	11.57 ± 0.18	87.2 ± 0.4
64	1.32 ± 0.04	9.87 ± 0.15	90.4 ± 0.3
128	$\mathbf{1.11} \pm 0.03$	$\mathbf{8.82} \pm 0.14$	$\mathbf{92.1} \pm 0.3$
256	1.29 ± 0.03	9.75 ± 0.13	91.3 ± 0.3

Concluding remark. As shown in Fig. 10, this experiment empirically characterizes the annotation-performance trade-off for MGR and provides guidance for practical deployment: choose K=128 that yields near-saturated accuracy (the "knee"), which minimizes labeling cost while retaining reconstruction and gaze performance.

C.4 Visualization of Gaze Consistency under Rotation

Motivation To further illustrate the effectiveness of our View-Consistent Regularization (VCR), we visualize predicted gaze vectors and eyeball projections under different synthetic rotations. This highlights how VCR enforces rotation-equivariant consistency in both appearance and structure.

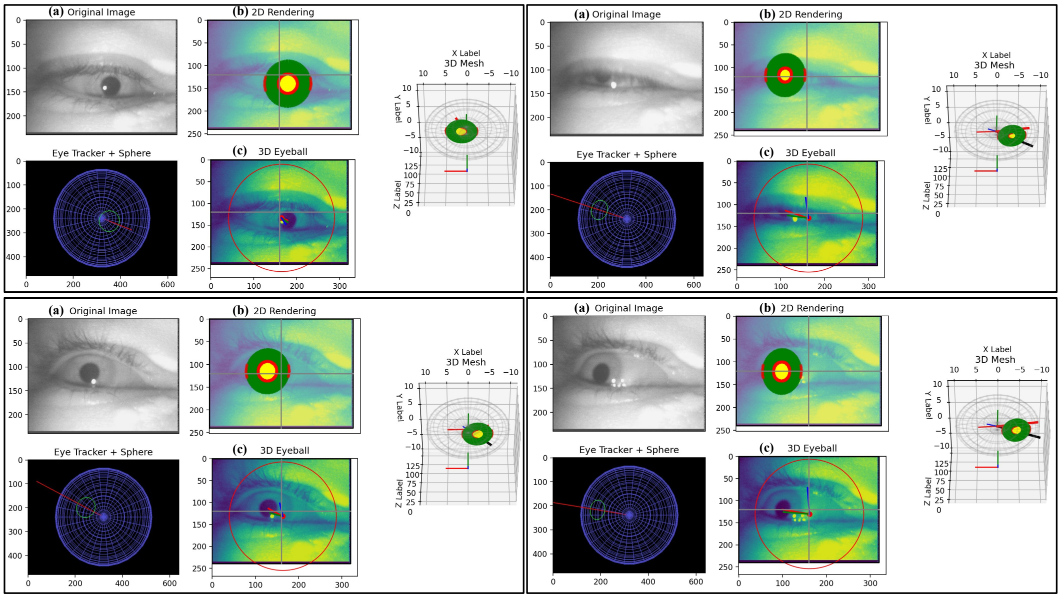

Figure 11: Visualization of gaze estimation under synthetic rotations. Each column shows (a) original image, (b) 2D rendering with VCR, and (c) rotated input with VCR. Red: ground-truth gaze; Blue: prediction without VCR; Green: prediction with VCR. With VCR, the predicted gaze aligns more consistently with the ground truth across different viewpoints, while the iris/pupil edges remain structurally faithful.

Visualization setup We apply yaw and pitch perturbations of $\pm 20^\circ$ and project the reconstructed eyeball structures before and after rotation. For each case, we plot:

- Ground-truth gaze vector (red arrow).
- Predicted gaze vector without VCR (blue arrow).
- Predicted gaze vector with VCR (green arrow).
- Corresponding 2D iris/pupil edge projections (overlayed in the image).

Discussion As shown in Fig. 11, models trained without VCR are sensitive to viewpoint changes, causing gaze vectors to drift away from the ground truth and inconsistent iris contours. In contrast, VCR enforces consistent predictions under rotations, leading to both geometrically faithful eyeball reconstructions and improved cross-view gaze alignment.

D Limitation and Future Work

D.1 Limitations

Although SG-Gaze demonstrates strong accuracy and cross-domain generalization, several limitations remain that open promising directions for future work.

(1) Subject-specific anatomical variation Our framework currently approximates gaze by aligning the optical axis with the visual axis and does not explicitly model subject-dependent kappa angle offsets. As shown in our supplementary experiments, this simplification may introduce residual bias across individuals. Future work will investigate lightweight calibration strategies or personalized modules to better adapt to subject-specific anatomy.

(2) Dependence on sparse 2D edge supervision The Model-Guided Reconstruction (MGR) branch relies on weak 2D edge labels (pupil and iris contours). While our sensitivity analysis shows that even sparse labels are effective, annotation effort is still required. Extending MGR to leverage unsupervised geometric cues or self-supervised contour discovery could further reduce dependence on human annotation.

(3) Synthetic-to-real domain gaps Although the proposed View-Consistent Regularization (VCR) alleviates domain gaps, our training pipeline still relies on synthetic perturbations that may not capture the full diversity of real-world conditions (e.g., extreme illumination, occlusions, eyeglasses). Incorporating physics-based rendering, domain adaptation, or generative data augmentation may further improve robustness.

(4) Deployment constraints Our method has not yet been fully optimized for resource-constrained AR/VR headsets. Exploring lightweight backbones, pruning, or distillation will be essential to enable real-time gaze tracking on mobile or embedded platforms.

D.2 Future Work

Building on SG-Gaze, several promising research directions can further align with the broader goals of the ICLR community:

Subject-adaptive representation learning: Develop lightweight calibration modules or meta-learning strategies to account for kappa angle offsets and anatomical variations, advancing personalized yet generalizable models.

Self-supervised structural learning: Move beyond annotated contours by exploiting self-supervised objectives and geometric consistency priors, enabling scalable training on large unlabeled datasets.

Physics-aware domain adaptation: Combine physically-grounded rendering, adversarial domain alignment, and generative augmentation to capture real-world shifts in illumination, occlusion, and device-specific imaging.

Resource-efficient model design: Pursue pruning, quantization, and distillation for low-latency deployment on AR/VR devices, bridging the gap between algorithmic advances and practical on-device applications.

Multi-task and cross-modal extensions: Explore joint learning with related tasks (e.g., iris recognition, user identification, eye-movement behavior analysis), and investigate integration with multi-modal signals such as speech or head motion for richer human-centric modeling.