GAZEMIXUP: DIRECTION-MAGNITUDE AWARE FEATURE INTERPOLATION FOR CROSS-DOMAIN GAZE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Appearance-based gaze estimation often suffers from poor cross-domain generalization due to variations in identity, illumination, imaging devices, and other domain-specific factors. We introduce GazeMixup, a geometry-aware feature augmentation method for cross-domain gaze estimation. The key idea is to decompose gaze features into direction and magnitude components, motivated by the observed association between feature magnitude and prediction error. GazeMixup interpolates these components separately with magnitude-aware coefficients and applies the same interpolation principle independently to the corresponding gaze labels, maintaining consistency between synthesized features and their supervision. The method is parameter-free, requires no target-domain data, and can be directly integrated into existing gaze estimation networks. Extensive experiments across multiple cross-domain tasks and backbone architectures show that GazeMixup consistently outperforms vanilla Mixup and improves baseline generalization on unseen domains, demonstrating the effectiveness of magnitude-aware feature interpolation for cross-domain gaze estimation.

1 INTRODUCTION

Gaze estimation aims to predict visual direction from images and has been widely studied for human–computer interaction (HCI) Wang et al. (2021), augmented and virtual reality (AR/VR) Liwan et al. (2025), driver monitoring Hu et al. (2025a), and attention analysis Varghese et al. (2024). Appearance-based methods, particularly CNN-based approaches, enable calibration-free gaze tracking in natural settings Vuillecard & Odobez (2025) and support saliency and scene understanding Kasahara et al. (2022); Kadner et al. (2023). Despite strong performance on established benchmarks, their accuracy can degrade on unseen datasets such as MPIIGaze Zhang et al. (2017b) and ETH-XGaze Zhang et al. (2020b) because of variations in subject appearance, illumination, imaging devices, and acquisition conditions Wang et al. (2022).

This challenge is commonly formulated as cross-domain generalization, where distribution shifts between source and target domains limit the transferability of learned gaze representations Xia et al. (2025); Bao & Lu (2024). Domain adaptation methods alleviate this problem by exploiting target-domain data, such as rotation-based alignment Liu et al. (2025) and contrastive regression with pseudo-labels Wang et al. (2019). However, their dependence on target samples limits their use when such data are unavailable. Domain generalization (DG) instead relies only on source-domain data. For example, feature purification attempts to suppress gaze-irrelevant factors in learned representations Cheng et al. (2022). Nevertheless, gaze-related and domain-dependent variations are often entangled in appearance-based features, making explicit separation difficult.

Data augmentation provides an alternative means of improving DG without accessing target-domain data. Mixup Zhang et al. (2017a) generates additional training samples through interpolation and has shown improved robustness in various tasks Wang et al. (2024). However, directly applying conventional Mixup to gaze regression is nontrivial because gaze labels represent viewing directions and do not necessarily exhibit the same Euclidean interpolation behavior as generic regression targets. Existing interpolation-based variants Zeng et al. (2024); Chen et al. (2022); Yun et al. (2019) in-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

troduce additional structures or assumptions, but do not explicitly consider the geometry of learned gaze features.

We therefore revisit the structure of gaze representations. A learned feature $\mathbf{f}$ can be decomposed into magnitude and direction as $\mathbf{f} = \|\mathbf{f}\|_2\mathbf{u}$, where $\mathbf{u}$ is the normalized direction. Our feature analysis shows that feature magnitude is associated with cross-domain prediction error, whereas feature direction has a stronger relationship with gaze variation. This observation suggests that treating the feature as a single Euclidean vector may overlook the different roles of these two components. Based on this insight, we propose GazeMixup, a geometry-aware feature augmentation method that separately interpolates feature direction and magnitude using magnitude-aware coefficients. The same interpolation principle is independently applied to the corresponding gaze labels, providing consistent feature–label construction.

GazeMixup is parameter-free and requires neither target-domain data nor additional networks, making it readily applicable to existing gaze estimation architectures. Rather than introducing an explicit disentanglement module, it modifies the interpolation process in the learned feature space. Experiments across multiple cross-domain tasks and backbone architectures show that GazeMixup improves over vanilla Mixup across the evaluated settings and achieves competitive performance with existing DG methods. Further analyses of feature distributions, error patterns, and label interpolation demonstrate the contribution of both feature geometry and feature–label consistent interpolation. The main contributions of this paper are summarized as follows:

- We propose GazeMixup, a parameter-free feature augmentation method that separately interpolates the magnitude and direction of gaze features without requiring target-domain data or additional networks.
- We identify the relationship between feature magnitude, directional variation, and cross-domain prediction error, and use these observations to motivate magnitude-aware feature interpolation.
- Extensive cross-domain experiments and ablation studies demonstrate the effectiveness of GazeMixup and verify the importance of jointly modeling feature direction, magnitude, and label interpolation.

2 MOTIVATION

Robust gaze estimation requires generalization across subjects, devices, and imaging conditions without target-domain data. This motivates source-only approaches that improve the transferability of learned gaze representations.

As shown in Fig. 1(a), a ResNet18 trained on a source domain exhibits substantial cross-domain error and feature-magnitude variation. After Mixup, both are reduced, with similar trends observed in Fig. 2, suggesting a relationship between feature magnitude and cross-domain behavior. We therefore decompose each feature into direction and magnitude. As shown in Fig. 1(b), magnitude is associated with angular prediction error, while direction shows a stronger correspondence with gaze variation, indicating that the two components should be treated differently during interpolation.

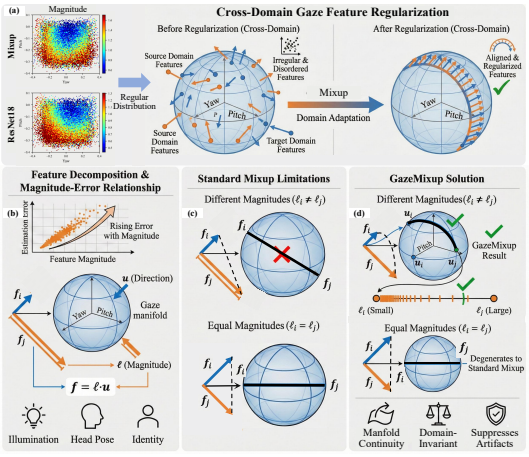

Figure 1: Schematic illustration of motivation.

This also exposes a limitation of standard Mixup. As illustrated in Fig. 1(c), uniform interpolation treats the feature as a whole and ignores the different roles of direction and magnitude, which may lead to undesirable magnitude variations in the generated features.

Based on these observations, we propose GazeMixup, which separately interpolates feature direction and magnitude. As illustrated in Fig. 1(d), directional interpolation preserves gaze-related variation, while magnitude-aware interpolation regulates feature scaling. The corresponding gaze labels follow the same magnitude-aware interpolation principle, providing consistent feature–label construction with a simple and architecture-agnostic augmentation strategy.

3 Related Work

3.1 Appearance-based gaze estimation

Appearance-based gaze estimation predicts gaze directly from images using CNNs on eye crops or full faces Cheng et al. (2024), sometimes incorporating additional cues Li et al. (2024); Zhao et al. (2024). Early datasets such as MPIIGaze Zhang et al. (2017b) (in-the-wild) and EYEDIAP Funes Mora et al. (2014) (controlled RGB/RGB-D) laid the foundation, while larger benchmarks like Gaze360 Kellnhofer et al. (2019) and ETH-XGaze Zhang et al. (2020b) increase diversity in pose, illumination, and viewpoint, highlighting cross-domain gaps that motivate this work.

To address domain shift, two main strategies exist: (i) unsupervised/semi-supervised domain adaptation using (often unlabeled) target data, and (ii) source-only domain generalization for unseen domains. Adversarial alignment and domain-adversarial training are widely used in adaptation Cheng et al. (2022); Kim et al. (2024); Wang et al. (2019); Tzeng et al. (2017), along with methods based on consistency, pseudo-labeling, or ensembles Liu et al. (2024); Guo et al. (2020), but they require target data or extra deployment steps.

In contrast, source-only methods improve robustness without target access, e.g., feature purification and causal/deconfounding strategies Kim et al. (2025); Liang et al. (2024). Recent work emphasizes feature geometry: projecting features onto structured spaces improves generalization Bao & Lu (2024), and suppressing harmful high-frequency components stabilizes performance Liu et al. (2025). Our approach follows this line by regularizing the feature manifold during source training via a geometry-aware Mixup that explicitly accounts for magnitude effects.

3.2 Mixup in representation learning

Mixup, originally designed for classification, has been extended to regression, where naive linear interpolation can produce unrealistic labels Zhang et al. (2017a). One line of work improves sampling and weighting by favoring local or label-consistent pairs to avoid harmful interpolation Hwang & Whang (2021); Yao et al. (2022); Hwang et al. (2024). Another line explores structure-aware strategies, such as cluster or anchor based perturbations, to better respect data geometry Schneider et al. (2023); Kan et al. (2023). The choice of interpolation path is also important, with curvature-aware methods improving generalization Kaufman & Azencot; 2024). These studies highlight the importance of interpolation design and motivate our focus on feature geometry and error distribution for gaze estimation.

In summary, prior work addresses cross-domain gaze estimation through dataset design, training strategies, architectures, and geometric modeling, often with added complexity. GazeMixup instead provides a simple, parameter-free feature-level augmentation tailored to the continuous and spherical nature of gaze, leveraging feature magnitude to improve source-only generalization.

4 Cross-domain Feature Analysis

We investigate the geometry of learned gaze features and its relationship with cross-domain prediction error. Given a source dataset $\mathcal{D}_s = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, a gaze estimator extracts

$$\mathbf{f}_i = F_\theta(\mathbf{x}_i), \qquad \hat{\mathbf{y}}_i = \mathrm{FC}(\mathbf{f}_i), \tag{1}$$

where $\mathbf{f}_i \in \mathbb{R}^n$ denotes the learned feature. Following Cheng et al. (2022), the feature may contain both gaze-related and domain-dependent variations, which motivates us to examine its internal geometry.

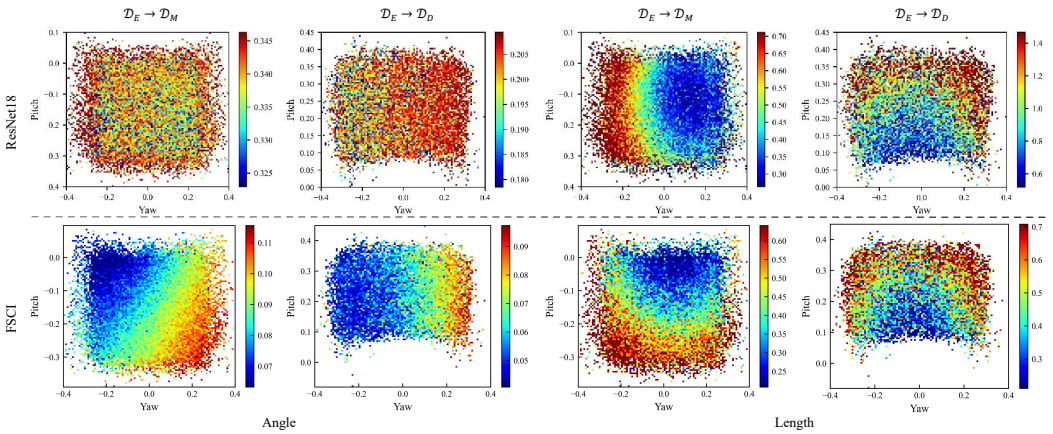

Figure 2: Visualization comparison between baseline and cross-domain models.

4.1 Feature Analysis

We decompose each feature into magnitude and direction:

$$\mathbf{f}_i = \ell_i \mathbf{u}_i, \qquad \ell_i = \|\mathbf{f}_i\|_2, \qquad \mathbf{u}_i = \frac{\mathbf{f}_i}{\|\mathbf{f}_i\|_2} \in \mathbb{S}^{n-1}. \tag{2}$$

Here, ℓ_i denotes the feature magnitude and $\mathbf{u}_i$ its normalized direction. This decomposition is used for analysis and does not impose a strict spherical constraint on the feature space, while being consistent with the geometric characteristics reported in Liu et al. (2017); Bao & Lu (2024).

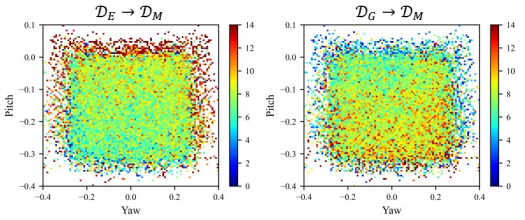

Figure 3: Variation of mean gaze error with gaze angle.

Cross-domain observations. We train ResNet18 on ETH-XGaze ($\mathcal{D}_E$) and Gaze360 ($\mathcal{D}_G$), and evaluate it on MPIIFaceGaze ($\mathcal{D}_M$). As shown in Fig. 3, the two source domains exhibit different error patterns on the same target domain, indicating different sensitivities to target-domain gaze variations. This motivates further analysis of the feature representations associated with these errors.

Magnitude and direction statistics. We analyze ℓ_i across gaze-angle bins and characterize $\mathbf{u}_i$ using a one-dimensional angular projection with $\mathbf{v} = e_1 = [1, 0, \ldots, 0]^\top$:

$$\psi_i = \arccos(\mathbf{u}_i^\top \mathbf{v}). \tag{3}$$

As shown in Fig. 2, feature magnitude varies non-uniformly across gaze angles, while models with stronger cross-domain capability, such as FSCI, exhibit more organized directional variation. These observations suggest that feature magnitude and direction have different relationships with cross-domain prediction.

Error–feature coupling. Figure 4 further compares prediction error with the two components. Prediction error tends to increase with feature magnitude, whereas its relationship with the one-dimensional directional projection is weaker. Thus, feature magnitude contains error-related information and provides an empirical basis for explicitly regulating feature magnitude during augmentation, without assuming that it corresponds to a specific nuisance factor.

5 Method

Based on these observations, we propose GazeMixup, which separately interpolates feature direction and magnitude and applies the same magnitude-aware interpolation principle to the corresponding gaze labels.

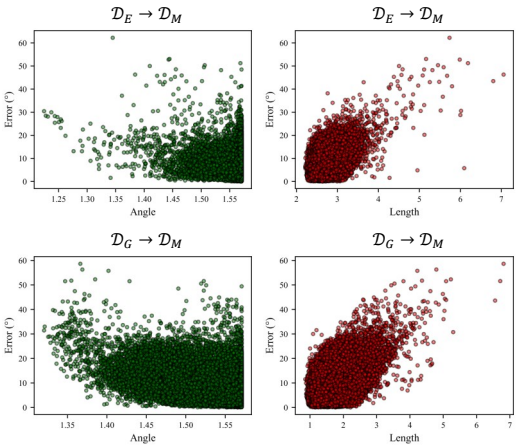

Figure 4: Relationship between mean gaze error and feature magnitude/directional components. ResNet18 is used as the baseline model, trained on $\mathcal{D}_E$ and $\mathcal{D}_G$, respectively.

5.1 MIXUP

Mixup Zhang et al. (2017a) generates virtual samples by linear interpolation:

$$\widetilde{\mathbf{x}} = \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j, \tag{4}$$

$$\widetilde{\mathbf{y}} = \lambda \mathbf{y}_i + (1-\lambda)\mathbf{y}_j, \tag{5}$$

where $\lambda \in [0,1]$ is sampled from a Beta distribution. However, conventional Mixup treats the feature vector as a whole and does not distinguish the roles of feature magnitude and direction.

5.2 GAZEMIXUP

For two randomly selected features, $\mathbf{f}_i = \ell_i \mathbf{u}_i$ and $\mathbf{f}_j = \ell_j \mathbf{u}_j$, GazeMixup separately interpolates their direction and magnitude:

$$\mathbf{f}_{\text{mix}} = \left(\lambda \frac{\mathbf{f}_i}{\|\mathbf{f}_i\|_2} + (1-\lambda)\frac{\mathbf{f}_j}{\|\mathbf{f}_j\|_2}\right)\left(\lambda\|\mathbf{f}_i\|_2 + (1-\lambda)\|\mathbf{f}_j\|_2\right). \tag{6}$$

Equivalently,

$$\mathbf{f}_{\text{mix}} = \alpha_f \lambda \mathbf{f}_i + \beta_f (1-\lambda)\mathbf{f}_j, \tag{7}$$

where

$$\alpha_f = \lambda + (1-\lambda)\frac{\|\mathbf{f}_j\|_2}{\|\mathbf{f}_i\|_2}, \qquad \beta_f = (1-\lambda) + \lambda\frac{\|\mathbf{f}_i\|_2}{\|\mathbf{f}_j\|_2}. \tag{8}$$

When the two feature magnitudes are equal, $\alpha_f = \beta_f = 1$ and GazeMixup reduces to conventional Mixup; otherwise, the interpolation is shifted toward the feature with smaller magnitude.

To maintain the same construction principle between features and supervision, we apply the magnitude-aware interpolation independently to the corresponding gaze labels:

$$\mathbf{y}_{\text{mix}} = \left(\lambda \frac{\mathbf{y}_i}{\|\mathbf{y}_i\|_2} + (1-\lambda)\frac{\mathbf{y}_j}{\|\mathbf{y}_j\|_2}\right)\left(\lambda\|\mathbf{y}_i\|_2 + (1-\lambda)\|\mathbf{y}_j\|_2\right), \tag{9}$$

or equivalently,

$$\mathbf{y}_{\text{mix}} = \alpha_y \lambda \mathbf{y}_i + \beta_y (1-\lambda)\mathbf{y}_j, \tag{10}$$

where

$$\alpha_y = \lambda + (1-\lambda)\frac{\|\mathbf{y}_j\|_2}{\|\mathbf{y}_i\|_2}, \qquad \beta_y = (1-\lambda) + \lambda\frac{\|\mathbf{y}_i\|_2}{\|\mathbf{y}_j\|_2}. \tag{11}$$

Here, (α_f, β_f) and (α_y, β_y) are computed independently from the feature and label magnitudes, respectively. Thus, the two spaces share the same magnitude-aware interpolation principle but do not share interpolation coefficients.

5.3 OPTIMIZATION

The synthesized feature–label pairs are optimized using an L_1 regression objective:

$$\theta^* = \arg\min_{\theta \in \Theta} \frac{1}{|M|} \sum_{i=1}^{M} \ell_{\text{L1}}\left(\text{FC}(\mathbf{f}_{\text{mix}}), \mathbf{y}_{\text{mix}}\right). \tag{12}$$

GazeMixup introduces no additional trainable parameters and requires neither target-domain data nor auxiliary supervision. Further analysis of the interpolation geometry is provided in Appendix A.

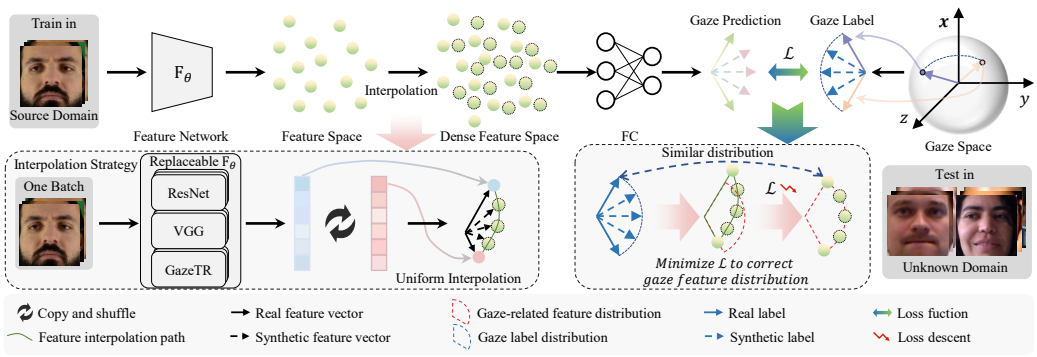

Figure 5: Framework of the proposed GazeMixup. By constructing nonlinear interpolation paths between arbitrary sample pairs, GazeMixup generates dense gaze representations and supervises learning through mixed gaze labels to capture the manifold distribution between samples.

Table 1: Cross-domain gaze estimation performance (°) on four domain transfer settings.

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
VGG	9.50°	19.14°	14.61°	19.94°
VGG+Mixup	6.72°	10.55°	7.69°	7.85°
VGG+GazeMixup	5.94°	8.57°	7.17°	7.74°
ResNet18	8.64°	7.83°	8.68°	12.35°
ResNet18+Mixup	7.45°	9.28°	7.37°	8.37°
ResNet18+GazeMixup	5.98°	7.67°	7.23°	7.94°
ResNet50	7.59°	8.90°	8.76°	11.76°
ResNet50+Mixup	7.42°	7.61°	7.11°	7.36°
ResNet50+GazeMixup	6.18°	7.45°	6.81°	7.10°
GazeTR	10.13°	12.08°	11.93°	12.87°
GazeTR+Mixup	7.06°	9.73°	8.52°	8.43°
GazeTR+GazeMixup	6.61°	8.29°	7.48°	8.08°

6 Experiments

6.1 Datasets

We evaluate our method on four widely adopted gaze estimation benchmarks: ETH-XGaze ($\mathcal{D}_E$) Zhang et al. (2020b), Gaze360 ($\mathcal{D}_G$) Kellnhofer et al. (2019), MPIIFaceGaze ($\mathcal{D}_M$) Zhang et al. (2017b), and EyeDiap ($\mathcal{D}_D$) Funes Mora et al. (2014). All datasets are preprocessed and normalized following the protocol in Zhang et al. (2017b) to ensure consistent input representation.

6.2 Implementation Details

All experiments are conducted using the GazeMixup framework implemented in PyTorch. The backbone network is initialized with weights pretrained on the ImageNet dataset Deng et al. (2009), while the remaining layers, including the feature fusion and regression heads, are initialized with default random values. We use a batch size of 128 and train the model for 80 epochs using the Adam optimizer. The initial learning rate is set to 1×10^{-4} and is scheduled using a cosine annealing strategy.

6.3 Results

Evaluation of GazeMixup. Table 1 reports cross-domain results under four transfer settings. GazeMixup consistently improves both the baseline and vanilla Mixup across all evaluated backbones. For example, on $\mathcal{D}_E \rightarrow \mathcal{D}_M$, the error decreases from 6.72° to 5.94° with VGG and from 7.45° to 5.98° with ResNet18. Similar improvements are observed under the other transfer settings,

Table 2: Comparison with state-of-the-art cross-domain gaze estimation methods and evaluation of error stability (ES). We report mean angular errors (in degrees) under four transfer settings and the proposed ES metric. "Target" indicates whether images from the target domain are required for training, and "Label" denotes whether additional annotations beyond gaze direction are used.

Method	Target	Label	$\mathcal{D}_E \to \mathcal{D}_M \downarrow$	$\mathcal{D}_E \to \mathcal{D}_D \downarrow$	$\mathcal{D}_G \to \mathcal{D}_M \downarrow$	$\mathcal{D}_G \to \mathcal{D}_D \downarrow$	ES ($\mathcal{D}_M$) $\downarrow$	ES ($\mathcal{D}_D$) $\downarrow$
Full-Face Zhang et al. (2017c)	×	×	12.35°	20.15°	11.13°	14.42°	1.22°	5.73°
PureGaze Cheng et al. (2022)	×	×	7.08°	7.48°	9.28°	9.32°	2.2°	1.84°
GazeAdv Wang et al. (2019)	✓	×	6.75°	8.10°	8.19°	12.27°	1.44°	4.17°
DAGEN Guo et al. (2020)	✓	×	6.16°	9.73°	7.61°	12.90°	1.45°	3.17°
ADDA Tzeng et al. (2017)	✓	×	6.33°	7.90°	8.76°	14.80°	2.43°	6.90°
GVBGD Cui et al. (2020)	✓	×	6.68°	7.27°	7.64°	12.44°	0.96°	5.17°
PnP-GA Liu et al. (2021)	✓	×	5.53°	5.87°	6.18°	7.92°	0.65°	2.05°
FSCI Liang et al. (2024)	×	✓	5.79°	6.96°	7.06°	7.99°	1.27°	1.03°
ResNet18+AGG Bao & Lu (2024)	×	×	7.10°	7.07°	7.87°	7.93°	0.77°	0.86°
ResNet50+AGG Bao & Lu (2024)	×	×	5.91°	6.75°	9.20°	11.36°	3.29°	4.61°
ResNet18+GazeMixup	×	×	5.98°	7.67°	7.23°	7.94°	1.25°	0.27°
ResNet50+GazeMixup	×	×	6.18°	7.45°	6.81°	7.10°	0.63°	0.35°

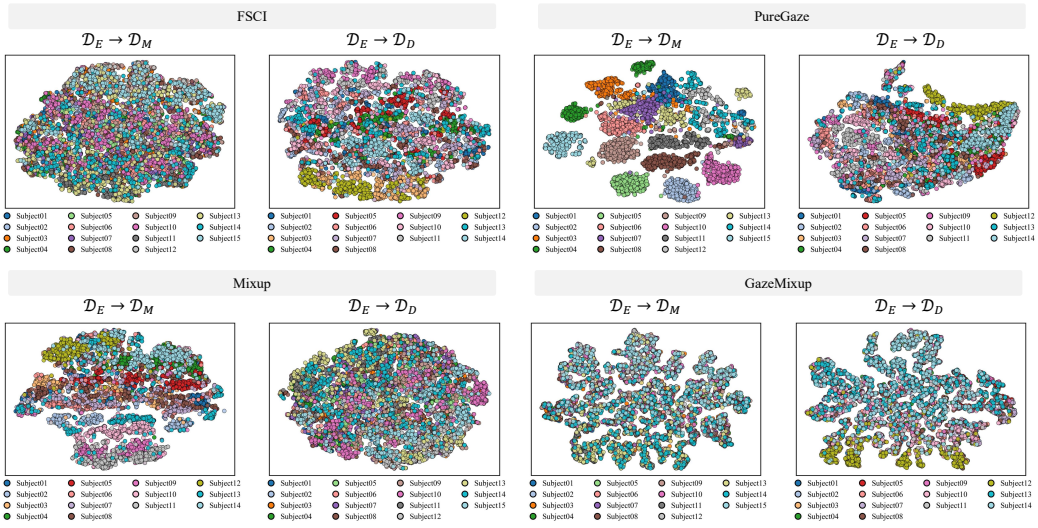

Figure 6: t-SNE visualization of gaze features colored by subject identity. Following PureGaze, samples with similar gaze directions are expected to exhibit less subject-specific clustering, while samples from the same subject should cover a broader range of gaze directions.

including the more challenging $\mathcal{D}_D$ target domain. More importantly, the gains are obtained without modifying the backbone or introducing target-domain data, indicating that the improvement comes from the interpolation strategy itself. Since the proposed interpolation explicitly separates feature direction and magnitude, these results support the use of feature geometry as a useful factor for constructing cross-domain training samples.

Comparison with SOTA Methods. Table 2 compares GazeMixup with representative cross-domain gaze estimation methods and reports the proposed Error Stability (ES) metric. GazeAdv Wang et al. (2019), DAGEN Guo et al. (2020), and ADDA Tzeng et al. (2017) rely on adversarial alignment or target-domain information, while FSCI Liang et al. (2024) uses auxiliary labels and AGG Bao & Lu (2024) introduces geometric constraints. GazeMixup instead operates only on source-domain training samples and does not introduce additional supervision. Its competitive cross-domain accuracy together with low ES suggests that the proposed interpolation is less sensitive to the choice of source domain.

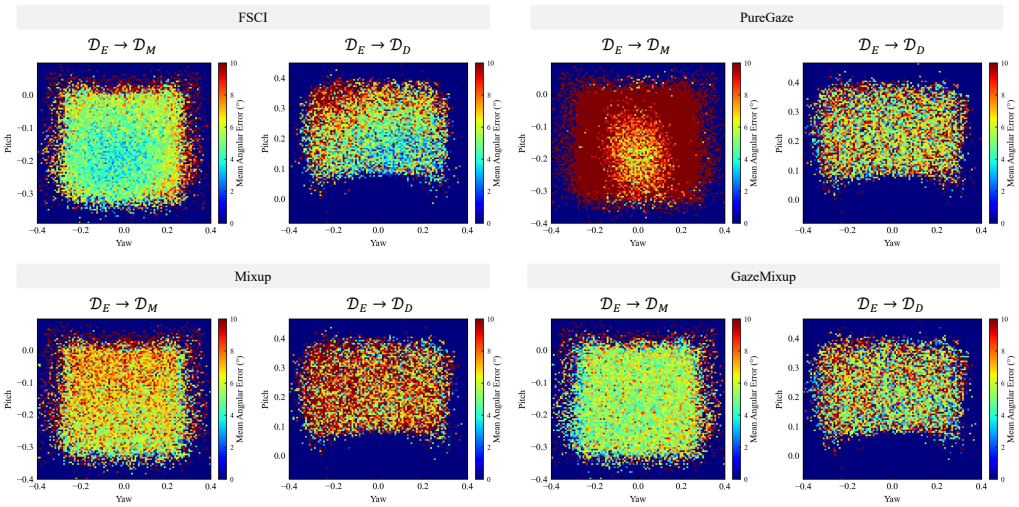

Figure 7: Uniform error distribution across gaze angles.

To quantify this property, let $\mathcal{E}_{A\to T}(i)$ and $\mathcal{E}_{B\to T}(i)$ denote the errors of sample i when models trained on source domains $\mathcal{D}_A$ and $\mathcal{D}_B$ are evaluated on the same target domain $\mathcal{D}_T$. We define

$$\mathrm{ES}(\mathcal{D}_T) = \frac{1}{N}\sum_{i=1}^{N} \left|\mathcal{E}_{A\to T}(i) - \mathcal{E}_{B\to T}(i)\right|, \tag{13}$$

where N is the number of target samples. Lower ES indicates that the prediction behavior is more consistent across source domains. GazeMixup achieves the lowest ES in Table 2, providing additional evidence that its gains are not limited to a particular source-domain configuration.

t-SNE Analysis of Extracted Gaze Features. We visualize the learned features using t-SNE and color the samples according to subject identity. Following Cheng et al. (2022), features of samples with similar gaze directions are expected to show limited subject-specific clustering. In Fig. 6, the baseline and vanilla Mixup exhibit clearer identity-related clusters, whereas GazeMixup produces a more interleaved distribution across subjects. The reduced clustering suggests that the learned representation is less organized around subject identity after training with GazeMixup. Combined with the cross-domain results, this finding provides qualitative evidence that the proposed interpolation changes the reliance of the model on subject-specific appearance.

Error Distribution Analysis. Figure 7 reports the mean gaze error over different gaze angles. FSCI and PureGaze show pronounced error fluctuations, particularly at extreme pitch and yaw angles. GazeMixup yields a noticeably flatter profile, with fewer large deviations across the angular range. The improvement is therefore distributed across different gaze configurations and is not driven by a small subset of angles. This behavior is consistent with the denser feature interpolation introduced by GazeMixup and complements the overall cross-domain results in Table 1.

Ablation Study on Feature Geometry. We compare GazeMixup with magnitude-only, direction-only, and normalized interpolation to examine the contribution of its two feature components. As reported in Table 3, GazeMixup achieves the lowest error in all four transfer settings. Direction-only interpolation performs better than magnitude-only interpolation, consistent with the stronger relationship between feature direction and gaze variation observed in Fig. 4. Still, retaining only direction is insufficient to reach the performance of GazeMixup, showing that feature magnitude also contributes useful information when incorporated with the directional component. Normalized interpolation performs poorly across the four settings, suggesting that discarding magnitude information altogether is not an effective way to regularize the feature representation.

Analysis of Feature Visualization. We project the extracted ResNet18 features into the length–angle space to examine their distribution after augmentation. As shown in Fig. 8, GazeMixup

Table 3: Ablation study of different feature interpolation strategies.

Method	DE→DM	DE→DD	DG→DM	DG→DD
Magnitude-only	8.35	8.92	9.11	10.28
Direction-only	7.13	8.38	8.16	8.63
Normalized Interpolation	8.71	9.16	8.57	9.17
GazeMixup	**5.98**	**7.67**	**7.23**	**7.94**

Table 4: Ablation study by changing only the label interpolation strategy while keeping the proposed feature interpolation fixed.

Label Interpolation	DE→DM	DE→DD	DG→DM	DG→DD
Convex Label Mixing	9.52	11.36	10.27	11.42
Normalized Label Interpolation	10.15	12.12	11.08	13.33
Magnitude-aware Label Interpolation	**5.98**	**7.67**	**7.23**	**7.94**

produces a clearer hemispherical magnitude pattern and smoother angular variation than standard Mixup. The resulting distribution is more organized across both dimensions, suggesting that separating feature direction and magnitude leads to more coherent feature variation. This observation is consistent with the magnitude-aware interpolation design and the improved cross-domain performance in Table 1.

Ablation Study on Label Interpolation. We further examine whether the label interpolation should follow the same magnitude-aware principle as the feature interpolation. We keep the proposed feature interpolation fixed and change only the label interpolation strategy. As shown in Table 4, both convex label mixing and normalized label interpolation result in substantially higher errors than the proposed magnitude-aware label interpolation across all four transfer settings. This comparison isolates the effect of the label construction and shows that the benefit of GazeMixup does not arise solely from feature interpolation. Instead, using consistent magnitude-aware interpolation principles in the feature and label spaces provides more effective supervision for the synthesized features.

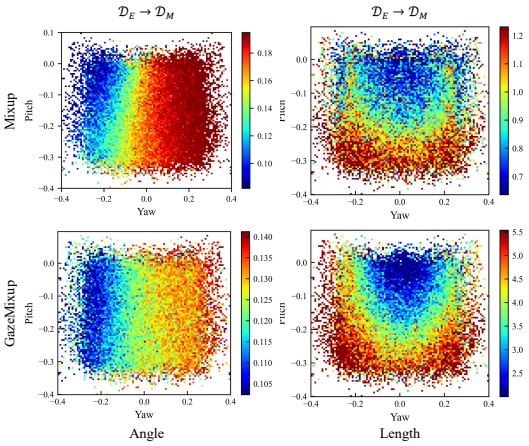

Figure 8: Visualization of extracted gaze feature distributions.

7 Conclusion

We proposed GazeMixup, a parameter-free geometry-aware feature augmentation method for cross-domain gaze estimation. By separately interpolating feature direction and magnitude with a magnitude-aware principle, GazeMixup improves the use of feature-level augmentation without requiring target-domain data or architectural changes. Extensive experiments and ablation studies across multiple datasets and backbones demonstrate its effectiveness over conventional Mixup and baseline models. Future work will investigate richer geometry-aware augmentation and its combination with unsupervised domain adaptation for larger domain shifts.

AI USE STATEMENT

Generative AI tools were used during the preparation of this manuscript for limited writing and programming support. In particular, AI-assisted tools were used to refine language, improve readability, and assist with software code development. All AI-assisted code was subsequently reviewed, tested, and validated by the authors. The authors also used language models to help identify grammatical issues and improve English expression. All scientific content, including the research questions, methodological design, experimental protocols, analysis, interpretation, and conclusions, was developed and verified by the authors. No generative AI tools were used to produce scientific figures, synthetic data, experimental results, or research ideas. All figures were created by the authors, with methodological diagrams and framework illustrations prepared as vector graphics, and result plots such as scatter plots exported at 600 dpi. Every AI-assisted output was reviewed and, where necessary, revised by the authors. The authors assume full responsibility for the accuracy, integrity, and final form of the manuscript, including its text, code, analyses, results, and figures.

REFERENCES

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.

Jie-Neng Chen, Shuyang Sun, Ju He, Philip HS Torr, Alan Yuille, and Song Bai. Transmix: Attend to mix for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12135–12144, 2022.

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, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7509–7528, 2024.

Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12455–12464, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 255–258, 2014.

Zidong Guo, Zejian Yuan, Chong Zhang, Wanchao Chi, Yonggen Ling, and Shenghao Zhang. Domain adaptation gaze estimation by embedding with prediction consistency. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Daosong Hu, Mingyue Cui, and Kai Huang. Fifa: Fine-grained inter-frame attention for driver's video gaze estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18760–18769, 2025a.

Daosong Hu, Mingyue Cui, and Kai Huang. Unmode: Uncertainty modeling for driver gaze estimation via feature disentanglement. *IEEE Transactions on Intelligent Transportation Systems*, 2025b.

Seong-Hyeon Hwang and Steven Euijong Whang. Regmix: Data mixing augmentation for regression. *arXiv preprint arXiv:2106.03374*, 2021.

Seong-Hyeon Hwang, Minsu Kim, and Steven Euijong Whang. Rc-mixup: A data augmentation strategy against noisy data for regression tasks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1155–1165, 2024.

Florian Kadner, Tobias Thomas, David Hoppe, and Constantin A Rothkopf. Improving saliency models' predictions of the next fixation with humans' intrinsic cost of gaze shifts. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pp. 2104–2114, 2023.

Xuan Kan, Zimu Li, Hejie Cui, Yue Yu, Ran Xu, Shaojun Yu, Zilong Zhang, Ying Guo, and Carl Yang. R-mixup: Riemannian mixup for biological networks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1073–1085, 2023.

Isaac Kasahara, Simon Stent, and Hyun Soo Park. Look both ways: Self-supervising driver gaze estimation and road scene saliency. In *European Conference on Computer Vision*, pp. 126–142. Springer, 2022.

Ilya Kaufman and Omri Azencot. Curvature enhanced data augmentation for regression. In *Forty-second International Conference on Machine Learning*.

Ilya Kaufman and Omri Azencot. First-order manifold data augmentation for regression learning. *arXiv preprint arXiv:2406.10914*, 2024.

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.

Suneung Kim, Woo-Jeoung Nam, and Seong-Whan Lee. Appearance debiased gaze estimation via stochastic subject-wise adversarial learning. *Pattern Recognition*, 152:110441, 2024.

Younghan Kim, Kangryun Moon, Yongjun Park, and Yonggyu Kim. Causal representation-based domain generalization on gaze estimation. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 386–395, 2025.

Nealson Li, Muya Chang, and Arijit Raychowdhury. E-gaze: Gaze estimation with event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4796–4811, 2024.

Ziyang Liang, Yiwei Bao, and Feng Lu. De-confounded gaze estimation. In *European Conference on Computer Vision*, pp. 219–235. Springer, 2024.

Ruicong Liu, Yunfei Liu, Haofei Wang, and Feng Lu. Pnp-ga+: Plug-and-play domain adaptation for gaze estimation using model variants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3707–3721, 2024.

Ruicong Liu, Haofei Wang, and Feng Lu. From gaze jitter to domain adaptation: Generalizing gaze estimation by manipulating high-frequency components. *International Journal of Computer Vision*, 133(3):1290–1305, 2025.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

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.

Lin Liwan, Wu Zongyu, Yijun Lu, Zhong Chen, and Guo Weijie. Recent progress on eye-tracking and gaze estimation for ar/vr applications: A review. *Electronics*, 14(17):3352, 2025.

Nora Schneider, Shirin Goshtasbpour, and Fernando Perez-Cruz. Anchor data augmentation. *Advances in Neural Information Processing Systems*, 36:74890–74902, 2023.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

Elizabeth B Varghese, Marwa Qaraqe, and Dena Al-Thani. Attention level evaluation in children with autism: leveraging head pose and gaze parameters from videos for educational intervention. *IEEE Transactions on Learning Technologies*, 17:1737–1753, 2024.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pp. 6438–6447. PMLR, 2019.

Pierre Vuillecard and Jean-Marc Odobez. Enhancing 3d gaze estimation in the wild using weak supervision with gaze following labels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13508–13518, 2025.

Kang Wang, Rui Zhao, Hui Su, and Qiang Ji. Generalizing eye tracking with bayesian adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11907–11916, 2019.

Xinming Wang, Jianhua Zhang, Hanlin Zhang, Shuwen Zhao, and Honghai Liu. Vision-based gaze estimation: A review. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):316–332, 2021.

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, 2022.

Zhicai Wang, Longhui Wei, Tan Wang, Heyu Chen, Yanbin Hao, Xiang Wang, Xiangnan He, and Qi Tian. Enhance image classification via inter-class image mixup with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17223–17233, 2024.

Lifan Xia, Yong Li, Xin Cai, Zhen Cui, Chunyan Xu, and Antoni B Chan. Collaborative contrastive learning for cross-domain gaze estimation. *Pattern Recognition*, 161:111244, 2025.

Huaxiu Yao, Yiping Wang, Linjun Zhang, James Y Zou, and Chelsea Finn. C-mixup: Improving generalization in regression. *Advances in neural information processing systems*, 35:3361–3376, 2022.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Zhichen Zeng, Ruizhong Qiu, Zhe Xu, Zhining Liu, Yuchen Yan, Tianxin Wei, Lei Ying, Jingrui He, and Hanghang Tong. Graph mixup on approximate gromov–wasserstein geodesics. In *Forty-first International Conference on Machine Learning*, 2024.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017a.

Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. How does mixup help with robustness and generalization? *arXiv preprint arXiv:2010.04819*, 2020a.

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017b.

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, 2017c.

Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 365–381. Springer, 2020b.

Guangrong Zhao, Yiran Shen, Chenlong Zhang, Zhaoxin Shen, Yuanfeng Zhou, and Hongkai Wen. Rgbe-gaze: A large-scale event-based multimodal dataset for high frequency remote gaze tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

A THEORY

Let $\mathbf{f} = \|\mathbf{f}\|_2 \, \mathbf{u} \in \mathbb{R}^d$ be the penultimate feature with magnitude $\|\mathbf{f}\|_2 \geq 0$ and direction $\mathbf{u} = \mathbf{f}/\|\mathbf{f}\|_2 \in \mathbb{S}^{d-1}$. Given two samples $(\mathbf{f}_i, \mathbf{y}_i)$ and $(\mathbf{f}_j, \mathbf{y}_j)$, draw $\lambda \sim \mathrm{U}(0,1)$ and define

$$\begin{aligned} \mathbf{u}_{\text{mix}} &= \lambda \, \mathbf{u}_i + (1-\lambda) \, \mathbf{u}_j, \\ \ell_{\text{mix}} &= \lambda \, \|\mathbf{f}_i\|_2 + (1-\lambda) \, \|\mathbf{f}_j\|_2, \\ \mathbf{f}_{\text{mix}} &= \mathbf{u}_{\text{mix}} \, \ell_{\text{mix}}, \end{aligned} \tag{14}$$

(no re-normalization of $\mathbf{u}_{\text{mix}}$). Let the inter-direction angle be $\theta = \arccos(\mathbf{u}_i^\top \mathbf{u}_j)$ and

$$\begin{aligned} c(\lambda, \theta) &= \big\| \lambda \, \mathbf{u}_i + (1-\lambda) \, \mathbf{u}_j \big\|_2 \\ &= \sqrt{\lambda^2 + (1-\lambda)^2 + 2\lambda(1-\lambda)\cos\theta}, \end{aligned} \tag{15}$$

so $\|\mathbf{f}_{\text{mix}}\|_2 = \ell_{\text{mix}} \cdot c(\lambda, \theta)$ and $\cos(\frac{\theta}{2}) \leq c(\lambda, \theta) \leq 1$. The gaze loss is L1:

$$\ell_{\text{L1}}\big(\hat{\mathbf{y}}, \mathbf{y}\big) = \big\| \hat{\mathbf{y}} - \mathbf{y} \big\|_1, \qquad \hat{\mathbf{y}} = \varphi(\mathbf{f}), \tag{16}$$

where φ is the regression head. We assume (*local*) Lipschitz continuity of $\ell_{\text{L1}} \circ \varphi$ in the magnitude and direction:

$$\begin{aligned} \big| \ell_{\text{L1}}(\varphi(l, u), \mathbf{y}) - \ell_{\text{L1}}(\varphi(l', u), \mathbf{y}) \big| &\leq L_\ell \, |l - l'|, \\ \big| \ell_{\text{L1}}(\varphi(l, u), \mathbf{y}) - \ell_{\text{L1}}(\varphi(l, u'), \mathbf{y}) \big| &\leq L_\theta \, \mathrm{dist}_{\mathbb{S}}(u, u'), \end{aligned} \tag{17}$$

where $\mathrm{dist}_{\mathbb{S}}$ is the geodesic distance on $\mathbb{S}^{d-1}$; $L_\ell, L_\theta > 0$ are Lipschitz constants.

Second-order statistics of $\mathbf{f}_{\text{mix}}$. Let $\bar{\ell} = (\|\mathbf{f}_i\|_2 + \|\mathbf{f}_j\|_2)/2$, $\Delta\ell = \|\mathbf{f}_i\|_2 - \|\mathbf{f}_j\|_2$, $\bar{\mathbf{u}} = (\mathbf{u}_i + \mathbf{u}_j)/2$, $\Delta\mathbf{u} = \mathbf{u}_i - \mathbf{u}_j$, and $s = \lambda - \frac{1}{2}$. With $\mathbb{E}[s] = 0$ and $\mathrm{Var}(s) = 1/12$ (uniform λ), a standard expansion around $\lambda = \frac{1}{2}$ gives

$$\begin{aligned} \mathbb{E}[\mathbf{f}_{\text{mix}}] &= \bar{\ell} \, \bar{\mathbf{u}} + \tfrac{1}{12} \, \Delta\ell \, \Delta\mathbf{u}, \\ \mathrm{Cov}(\mathbf{f}_{\text{mix}}) &\approx \tfrac{1}{12} \big(\Delta\ell \, \bar{\mathbf{u}} + \bar{\ell} \, \Delta\mathbf{u} \big) \big(\Delta\ell \, \bar{\mathbf{u}} + \bar{\ell} \, \Delta\mathbf{u} \big)^\top. \end{aligned} \tag{18}$$

Moreover,

$$\begin{aligned} \mathrm{Var}(\|\mathbf{f}_{\text{mix}}\|_2) &= \mathrm{Var}(\ell_{\text{mix}} \, c) \\ &\leq \mathrm{Var}(\ell_{\text{mix}}) \, \mathbb{E}[c^2] + \bar{\ell}^2 \, \mathrm{Var}(c), \end{aligned} \tag{19}$$

and

$$\begin{aligned} \mathbb{E}[c^2] &= \mathbb{E}\big[\lambda^2 + (1-\lambda)^2\big] + 2 \, \mathbb{E}\big[\lambda(1-\lambda)\big] \cos\theta \\ &= \tfrac{2}{3} + \tfrac{1}{3} \cos\theta. \end{aligned} \tag{20}$$

Eqs. equation 18–equation 20 are obtained by linearizing $\mathbf{f}_{\text{mix}}$ in s and using $\|\Delta\mathbf{u}\|_2^2 = 4\sin^2(\frac{\theta}{2})$, $\|\bar{\mathbf{u}}\|_2^2 = \cos^2(\frac{\theta}{2})$.

Adversarial loss and the upper bound. Define the (*feature-space*) L1 adversarial loss with perturbation budget $\varepsilon > 0$ by

$$\mathcal{L}_{\text{adv}}^{\text{L1}}(\varepsilon) = \mathbb{E}\left[\sup_{\|\boldsymbol{\delta}\|_2 \leq \varepsilon} \ell_{\text{L1}}\big(\varphi(\mathbf{f}_{\text{mix}} + \boldsymbol{\delta}), \mathbf{y}_{\text{mix}}\big) \right]. \tag{21}$$

Following the mixup analysis template of Zhang et al. (2020a) (Theorem 1) and the hidden-state interpolation idea of Verma et al. (2019), we take a second-order Taylor expansion of $\ell_{\text{L1}} \circ \varphi$ at the mean $\bar{\mathbf{f}} = \mathbb{E}[\mathbf{f}_{\text{mix}}]$:

$$\ell_{\text{L1}}\big(\varphi(\bar{\mathbf{f}} + \mathbf{z}), \mathbf{y}\big) \approx \ell_{\text{L1}}\big(\varphi(\bar{\mathbf{f}}), \mathbf{y}\big) + \mathbf{g}^\top \mathbf{z} + \tfrac{1}{2} \, \mathbf{z}^\top \mathbf{H} \, \mathbf{z}, \tag{22}$$

where $\mathbf{g} = \nabla_{\mathbf{f}} (\ell_{\text{L1}} \circ \varphi)|_{\bar{\mathbf{f}}}$ and $\mathbf{H}$ is the local Hessian. Let $\mathbf{z} = (\mathbf{f}_{\text{mix}} - \bar{\mathbf{f}}) + \boldsymbol{\delta}$. Taking expectation and upper-bounding the linear and quadratic terms via Cauchy–Schwarz and spectral norms yields

$$\begin{aligned} \mathcal{L}_{\text{adv}}^{\text{L1}}(\varepsilon) \leq \mathcal{L}_{\text{ERM}}^{\text{L1}} &+ \|\mathbf{g}\|_2 \, \mathbb{E}\big[\|\mathbf{f}_{\text{mix}} - \bar{\mathbf{f}}\|_2\big] \\ &+ \tfrac{1}{2} \|\mathbf{H}\|_2 \, \mathbb{E}\big[\|\mathbf{f}_{\text{mix}} - \bar{\mathbf{f}}\|_2^2\big] + C \, \varepsilon^2. \end{aligned} \tag{23}$$

where $\mathcal{L}_{\text{ERM}}^{\text{L1}} = \mathbb{E}[\ell_{\text{L1}}(\varphi(\bar{\mathbf{f}}), \mathbf{y})]$ and C collects the $\boldsymbol{\delta}$-only terms (bounded by ε^2). To connect the Euclidean moments in equation 22 with the magnitude–direction geometry, we decompose the deviation into radial and angular parts and invoke the Lipschitz bounds in equation 17. This yields constants $K_1, K_2 > 0$ (depending on local curvature through $\|\mathbf{g}\|_2, \|\mathbf{H}\|_2$ and the perturbation budget) such that

$$\begin{aligned}\mathcal{L}_{\text{adv}}^{\text{L1}}(\varepsilon_{\text{mix}}) \le \mathcal{L}_{\text{ERM}}^{\text{L1}} + K_1\, L_\ell^2\, \mathrm{Var}(\|\mathbf{f}_{\text{mix}}\|_2) \\ +\; K_2\, L_\theta^2\, \mathbb{E}\Big[\mathrm{dist}_{\mathbb{S}}(\mathbf{u}_{\text{mix}}, \bar{\mathbf{u}})^2\Big]\,,\end{aligned} \tag{24}$$

where $\bar{\mathbf{u}} = \mathbb{E}[\mathbf{u}_{\text{mix}}]$ and the two moment terms are controlled by equation 19–equation 20 (hence by θ and λ). Eq. equation 24 is the feature-space analogue of the adversarial upper bound proved for input-space mixup in Zhang et al. (2020a); our bound inherits the *vicinal risk regularization* view of mixup and replaces the raw Euclidean moments by geometry-aware quantities: (i) a radial variance term $\mathrm{Var}(\|\mathbf{f}_{\text{mix}}\|_2)$ that is *shrunk* by $c(\lambda, \theta)$ when θ is large, and (ii) an angular fluctuation term $\mathbb{E}\big[\mathrm{dist}_{\mathbb{S}}(\mathbf{u}_{\text{mix}}, \bar{\mathbf{u}})^2\big]$ which is small in the small-angle regime. Since these statistics enter Eq. equation 24 additively with positive weights $K_1 L_\ell^2$ and $K_2 L_\theta^2$, any reduction in either term *strictly lowers* the adversarial L1 bound; accordingly, radial compression reduces inter-source magnitude mismatch while angular contraction smooths per-angle error histograms, yielding a smaller ES and improved cross-domain MAE.

Explanation of symbols. λ (mixing weight); θ (angle between $\mathbf{u}_i, \mathbf{u}_j$); $c(\lambda, \theta)$ (directional shrinkage factor); ℓ_{mix} (mixed magnitude); $\mathbf{f}_{\text{mix}}$ (mixed feature); $\bar{\ell}, \Delta\ell, \bar{\mathbf{u}}, \Delta\mathbf{u}$ (pairwise mean/difference); ℓ_{L1} (L1 gaze loss); φ (regression head); L_ℓ, L_θ (Lipschitz constants); $\mathrm{dist}_{\mathbb{S}}$ (geodesic distance on $\mathbb{S}^{d-1}$); $\mathcal{L}_{\text{ERM}}^{\text{L1}}$ (ERM objective with L1 loss); $\mathcal{L}_{\text{adv}}^{\text{L1}}$ (feature-space adversarial loss); $\mathbf{g}, \mathbf{H}$ (gradient/Hessian at $\bar{\mathbf{f}}$); K_1, K_2 (curvature- and budget-dependent positive constants); ε_{mix} (effective perturbation radius).

B Analysis of interpolation sample count.

We study the effect of interpolation density by varying the number of mixed samples for cross-domain gaze estimation. As shown in Fig. 9, GazeMixup consistently outperforms Mixup, indicating that nonlinear interpolation better captures the gaze manifold and improves stability. As the number of samples increases, GazeMixup's error gradually flattens, suggesting that the model has learned a stable gaze representation and additional samples mainly enhance smoothness.

We attribute this convergence to the fact that once the geometric relation between appearance and gaze is captured, remaining errors are dominated by physiological limits such as the offset between optical and visual axes. In contrast, Mixup shows more fluctuation due to limited sampling, pose imbalance, and illumination variation. GazeMixup provides smoother transitions and denser manifold coverage, resulting in more stable and robust performance across different interpolation densities.

C Analysis of feature dimension:

We set the feature dimension to 128 as it provides the best trade-off between representation capacity, interpolation effectiveness, and manifold density. As shown in Table 5, smaller dimensions (e.g., 64) degrade performance due to information compression and reduced discriminability, which weakens interpolation. Larger dimensions (e.g., 512 or 1024) also fail to improve performance. Higher-dimensional spaces reduce sample density, making interpolation less reliable, and increase optimization difficulty and overfitting risk. They also tend to preserve nuisance variations such as illumination and identity, which harms generalization. Thus, 128 forms a sweet spot. It is expressive enough to encode gaze structure while remaining compact for dense and stable interpolation. It also reduces computational cost. Both smaller and larger dimensions lead to worse cross-domain performance, so we adopt 128 as the default.

C.1 In-Domain Evaluation

Although GazeMixup is primarily designed for source-only cross-domain generalization, we further evaluate its effect under the in-domain setting. As shown in Table 6, GazeMixup consistently im-

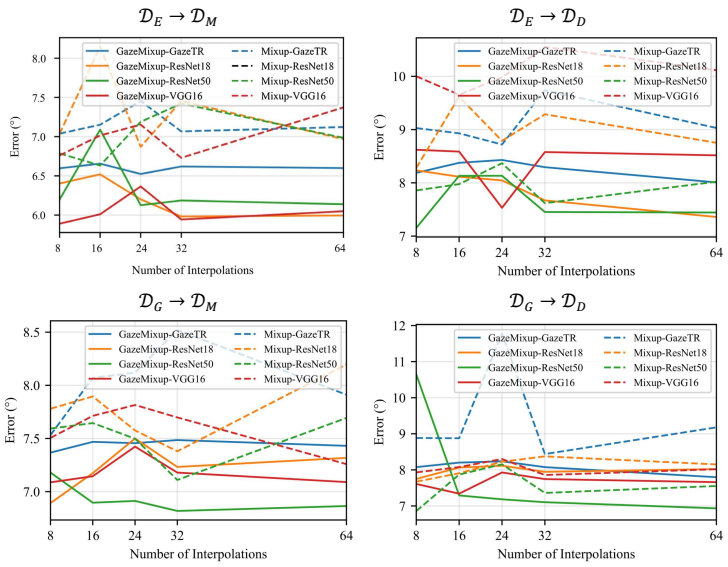

Figure 9: Effect of interpolation sample count on gaze estimation.

Table 5: Effect of feature dimensionality on cross-domain interpolation.

Dimension	Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
64	VGG	7.03°	8.91°	7.69°	8.53°
	ResNet18	6.49°	8.40°	7.35°	8.16°
	ResNet50	6.54°	8.13°	7.29°	7.38°
	GazeTR	7.10°	8.79°	7.42°	7.56°
128	VGG	**5.94°**	8.57°	7.17°	7.74°
	ResNet18	5.98°	7.67°	7.23°	7.94°
	ResNet50	6.18°	7.45°	**6.81°**	7.10°
	GazeTR	6.61°	8.29°	7.48°	8.07°
256	VGG	6.61°	8.28°	7.17°	7.85°
	ResNet18	6.43°	7.61°	7.42°	8.08°
	ResNet50	6.71°	**6.91°**	7.00°	7.48°
	GazeTR	6.51°	7.55°	7.26°	8.07°
512	VGG	6.14°	8.59°	7.64°	7.10°
	ResNet18	6.35°	8.20°	7.28°	7.27°
	ResNet50	6.70°	8.47°	7.60°	**6.95°**
	GazeTR	6.66°	9.87°	7.80°	7.48°
1024	VGG	6.52°	8.56°	7.28°	8.71°
	ResNet18	6.91°	7.72°	7.17°	8.18°
	ResNet50	6.62°	7.54°	7.46°	8.18°
	GazeTR	6.88°	8.16°	7.81°	7.91°

proves the baseline models on both ETH-XGaze and Gaze360, while conventional Mixup provides only limited improvements. This result suggests that geometry-aware interpolation not only improves cross-domain robustness, but also increases the local sampling density of the learned feature space and improves the learned gaze regression function.

D LIMITATIONS

While GazeMixup improves cross-domain generalization, it still has several limitations. As a source-only strategy, it does not utilize unlabeled target-domain data, unlike domain adaptation methods. Therefore, its performance may be constrained in scenarios where target data are available

Table 6: In-domain gaze estimation performance on ETH-XGaze and Gaze360.

Method	ResNet18		ResNet50		VGG	
	ETH-XGaze	Gaze360	ETH-XGaze	Gaze360	ETH-XGaze	Gaze360
Baseline	5.08	12.73	5.35	12.37	5.12	12.15
Mixup	5.12	12.56	5.18	12.46	5.11	12.37
AGG	5.56	13.03	6.29	15.63	5.78	13.13
GazeMixup	**4.93**	**11.78**	**5.02**	**11.45**	**5.09**	**11.24**

and could be exploited for further alignment. In addition, we observe that the performance gains tend to saturate. The remaining errors are likely caused by intrinsic physiological factors, such as the offset between the optical axis and the visual axis of the human eye, which cannot be corrected through feature interpolation alone. This suggests that there exists an upper bound for improvement when relying purely on augmentation-based strategies.

GazeMixup also assumes that the interpolated features lie close to valid samples on the underlying gaze manifold. However, this assumption may not always hold, especially in cases involving extreme gaze angles or highly sparse training data. In such situations, the generated samples may still carry residual identity or environmental biases, which can limit the effectiveness of the method. Finally, although GazeMixup reduces nuisance variations, it does not explicitly remove them. As a result, some domain-specific noise may remain in the learned representations. The overall effectiveness of the method therefore depends on the quality and geometry of the underlying feature space, and may vary across datasets and model architectures.