HalfMix Augmentation and Regularized Dual-Path Learning for Cross-Domain Gaze Estimation

Jiuk Hong
hong4497@knu.ac.kr
Heechul Jung
heechul@knu.ac.kr

Department of Artificial Intelligence
Kyungpook National University
Daegu, South Korea

Abstract

Cross-domain gaze estimation presents a persistent challenge in computer vision, as models often experience significant performance degradation when applied to unseen target domains with different characteristics (e.g., subjects, illumination, camera setups). To address this, we first introduce HalfMix augmentation, a novel data augmentation technique specifically designed for gaze estimation. HalfMix mitigates common issues in conventional mix-based augmentations like MixUp and CutMix by preserving crucial eye regions without overlap or occlusion. Secondly, we propose a regularized dual-path learning strategy to effectively capitalize on the rich, dual-gaze information inherent in HalfMix-generated samples. This strategy employs a dual-path architecture where a shared encoder feeds into two distinct prediction pathways. To foster diverse and robust feature learning, we incorporate two key regularization components: diversity-promoting regularization (DPR) and dual-gaze feature alignment (DGFA). Extensive experiments on several benchmark datasets demonstrate that our integrated approach significantly improves cross-domain gaze estimation performance, outperforming existing methods by learning more robust and generalizable gaze representations that are less sensitive to domain shifts. Code is available at https://github.com/CreamNuts/HalfMix-Dual-Path-Learning.

1 Introduction

Gaze estimation is pivotal for understanding human attention and intention, with wide-ranging applications in human-computer interaction, virtual reality, and driver assistance systems [2, 3, 9]. While appearance-based gaze estimation methods using deep learning have achieved considerable success [5], a significant challenge persists: their performance often degrades substantially when applied to unseen domains (i.e., cross-domain generalization) due to variations in subject appearance, illumination conditions, and camera setups. Consequently, enhancing cross-domain robustness is a critical research avenue in gaze estimation. Many studies [1, 4, 10, 11, 13, 14] have focused on suppressing gaze-irrelevant features to improve cross-domain performance, yet there is still room for improvement.

Data augmentation is a widely adopted strategy for enhancing domain generalization in various computer vision tasks. However, popular mix-based augmentation techniques such

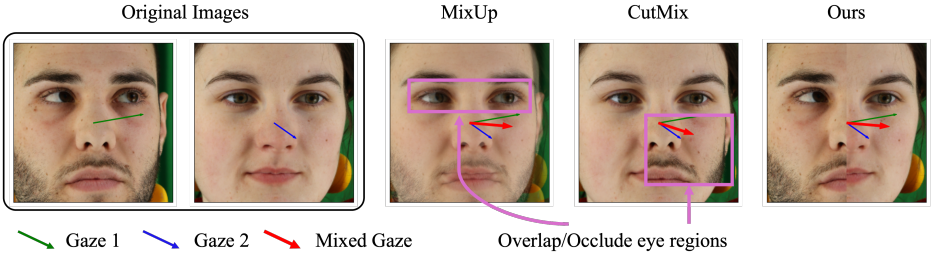

Figure 1: Overview of the proposed HalfMix augmentation, illustrating how it combines two source images while preserving crucial eye regions. This is contrasted with potential issues arising from other mixing strategies like MixUp (potential eye region overlap) and CutMix (potential eye region occlusion).

as MixUp [18] and CutMix [17] are not straightforwardly applicable to gaze estimation. This difficulty arises from two primary characteristics of gaze data: first, gaze labels are sensitive to geometric transformations, meaning that applying such transformations can easily corrupt the ground truth gaze direction; second, gaze estimation is fundamentally a regression problem, and the application of mix-based methods to regression tasks, while explored, is not as prevalent or well-established as in classification tasks. Indeed, as shown in our experiments (see Table 4), naive applications of MixUp or CutMix to gaze estimation do not yield significant improvements and, in some instances, even lead to performance degradation. Figure 1 illustrates potential issues: MixUp can create noisy samples by overlapping eye regions from two images, while CutMix, by pasting a random patch, might obscure the crucial eye regions, forcing the model to rely on gaze-irrelevant subject appearance cues.

To address these limitations, this paper introduces HalfMix augmentation, a novel data augmentation technique specifically designed for gaze estimation. As depicted in Figure 1, HalfMix combines halves of two source images, ensuring that the critical eye regions from both images are preserved and do not overlap. This approach mitigates the issues of information interference (as in MixUp) and feature omission (as in CutMix) while being simple to implement. Furthermore, to effectively leverage the rich information present in HalfMix-generated samples, which contain appearance and gaze information from two different subjects in a single image, we propose a novel learning strategy called regularized dual-path learning. This strategy involves a dual-path architecture where a shared feature encoder processes the HalfMix image, and its output is then fed into two separate prediction pathways, each tasked with predicting one of the original gaze labels. Simply employing two classifiers could lead to redundant feature learning or an effect similar to merely doubling the gradient magnitude, potentially causing overfitting on in-domain data. To counteract this, we introduce two regularization techniques: 1) diversity-promoting regularization (DPR), which encourages the two pathways to learn diverse features, and 2) dual-gaze feature alignment (DGFA), a refinement technique that encourages the feature representations corresponding to the two gaze labels within a HalfMix sample to be similar if the labels themselves are close.

Our contributions are threefold: 1) HalfMix augmentation, a simple yet effective data augmentation method tailored for gaze estimation. 2) Regularized dual-path learning, a novel training strategy incorporating DPR and DGFA to effectively learn from HalfMix data. and 3) Extensive experimental validation demonstrating that our proposed methods significantly

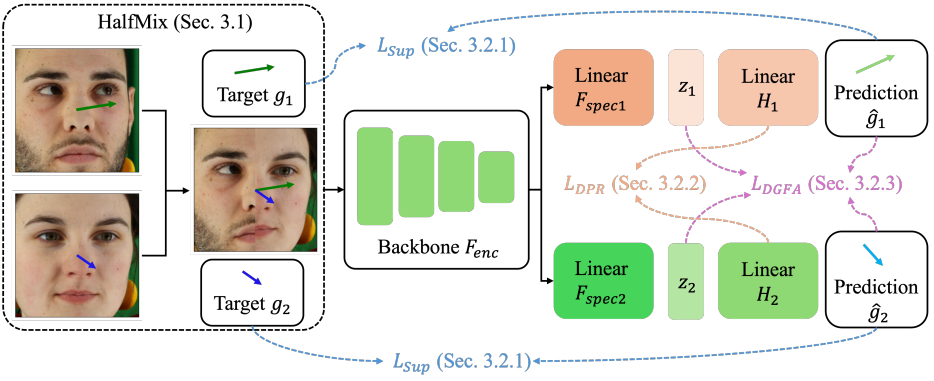

Figure 2: Overall framework of our proposed method, illustrating HalfMix augmentation and the regularized dual-path learning strategy, which includes a dual-path architecture, diversity-promoting regularization (DPR), and dual-gaze feature alignment (DGFA).

improve cross-domain gaze estimation performance by learning more robust and generalizable gaze representations.

2 Related Works

Gaze Estimation Strategies. Since most datasets [6, 8, 19, 21] for gaze estimation have different gaze directions and head pose distributions, cross-dataset gaze estimation that involves training and testing on different datasets is an important problem. Liu et al. [11] propose a domain adaptation framework, which is an ensemble of networks that learn collaboratively with the guidance of outliers. Cheng et al. [4] purify the gaze feature by designing a self adversarial framework. Wang et al. [14] modify the supervised contrastive loss to adapt regression task and use pseudo-label for domain adaptation. Note that these methods are based on the regression loss, then we can easily adapt our method to those methods.

Mix-based Data Augmentation for Gaze Estimation. Mix-based DA methods [12, 17, 18] provide strong regularization, improving both generalization and robustness to adversarial attacks in image classification. However, the distribution of linearly interpolated continuous labels is not similar as the true distribution of continuous labels, leading to degradation of generalization in both regression and ordinal regression. Hwang and Whang [7] try to solve this problem by training additional mixing controller based on reinforcement learning, and the method outperforms naive adaptation of MixUp-based augmentations. Yao et al. [16] propose a simple sampling algorithm where the closer the labels are, the greater the probability of being selected in MixUp.

3 Method

Our proposed method consists of two main contributions: HalfMix augmentation, a novel data augmentation technique tailored for gaze estimation, and regularized dual-path learning, a strategy to effectively learn from HalfMix-generated data. Figure 2 illustrates the overall framework.

3.1 HalfMix Augmentation

To address the limitations of conventional mix-based augmentations for gaze estimation, we propose HalfMix. Given two source images I_a and I_b with corresponding gaze labels g_a and g_b, randomly sampled from the current batch, HalfMix generates a composite image I_{mix} by combining halves of I_a and I_b. For instance, using a vertical split at the image center ($W/2$ for width W):

$$I_{mix}(x,y) = \begin{cases} I_a(x,y) & \text{if } y \leq W/2 \\ I_b(x,y) & \text{if } y > W/2 \end{cases}. \tag{1}$$

This ensures that crucial eye regions from both I_a and I_b are preserved and do not overlap in I_{mix}. The model is then trained to predict both original gaze labels, g_a and g_b, from the single composite image I_{mix}. This provides rich, dual supervisory signals from distinct gaze directions within one sample.

The key advantage of HalfMix lies in its preservation of gaze information, as each half maintains the complete eye region and its spatial relationship to facial landmarks. Unlike MixUp which alpha-blends pixels (corrupting geometric features) or CutMix which may randomly occlude critical eye regions, HalfMix keeps the eye-to-face geometry intact for both gaze vectors. No explicit constraints are imposed on head pose or gaze direction differences during pair selection, allowing the model to learn from diverse image combinations.

3.2 Regularized Dual-Path Learning

To effectively leverage the dual-gaze information from HalfMix images, we introduce a regularized dual-path learning strategy. This involves a specialized architecture and regularization techniques to promote diverse and robust feature learning.

3.2.1 Dual-Path Architecture

A HalfMix image I_{mix} (derived from I_a and I_b with labels g_a, g_b) is processed by a shared feature encoder F_{enc} to produce $f_{common} = F_{enc}(I_{mix})$. This common feature representation is then channeled into two separate pathways. Each pathway $k \in \{1,2\}$ consists of a path-specific feature transformation F_{spec_k} and a prediction head H_k, yielding two gaze predictions: $\hat{g}_1 = H_1(F_{spec_1}(f_{common}))$ and $\hat{g}_2 = H_2(F_{spec_2}(f_{common}))$. When HalfMix is not used, both pathways predict the same target ($g_a = g_b$). The primary supervised learning objective is to train these pathways to predict the original gaze labels. The total supervised loss L_{sup} is an average of the losses for each path, using a gaze estimation loss function $\mathcal{L}$ (e.g., L1 loss):

$$L_{sup} = \frac{1}{2}\left(\mathcal{L}(\hat{g}_1, g_a) + \mathcal{L}(\hat{g}_2, g_b)\right). \tag{2}$$

This gives equal importance to both original labels derived from the HalfMix augmentation.

During testing, we process standard single images without HalfMix augmentation. The input image is fed through the shared encoder F_{enc}, and both pathways generate predictions: $\hat{g}_1 = H_1(F_{spec_1}(f_{common}))$ and $\hat{g}_2 = H_2(F_{spec_2}(f_{common}))$. The final prediction is their average: $\hat{g}_{final} = (\hat{g}_1 + \hat{g}_2)/2$, leveraging the complementary features learned during training.

3.2.2 Diversity-Promoting Regularization (DPR)

To prevent the two pathways from learning redundant features and to encourage them to capture complementary aspects of the input, we apply diversity-promoting regularization

(DPR). Without explicit regularization, the weight matrices W_1 and W_2 rapidly converge to near-identical values (cosine similarity > 0.95 and KL divergence < 0.005 after 3 epochs), essentially collapsing to a single pathway with doubled gradient magnitude. This redundancy eliminates the benefits of dual-path learning. To address this, DPR incorporates two distinct terms. Let W_1 and W_2 represent the learnable parameters (e.g., weights of the final layer) of the prediction heads H_1 and H_2, respectively.

First, to ensure the pathways learn diverse representations, we apply a cosine similarity penalty, L_{cs}. This term directly penalizes high cosine similarity between W_1 and W_2, encouraging them to be less aligned:

$$L_{cs}(W_1, W_2) = \frac{W_1 \cdot W_2}{||W_1||_2 \cdot ||W_2||_2}. \tag{3}$$

Minimizing the contribution of this term (as part of L_{DPR} which is added to the total loss) effectively maximizes the dissimilarity between the weight vectors, promoting diversity.

Second, while diversity is crucial, we also found experimentally that allowing the weight distributions to become excessively dissimilar (i.e., a very large KL divergence) can negatively impact performance. Therefore, we introduce a regularization term L_{KL} based on the Kullback-Leibler (KL) divergence, applied to the weights W_1 and W_2 (after transforming them into probability distributions, e.g., via Softmax, denoted by $Q(\cdot)$). This term helps to keep the weight distributions from diverging too drastically, acting as a stabilizer:

$$L_{KL}(W_1, W_2) = \frac{1}{2}\left(D_{KL}(Q(W_1)||Q(W_2)) + D_{KL}(Q(W_2)||Q(W_1))\right), \tag{4}$$

where $D_{KL}(\cdot||\cdot)$ denotes the KL divergence. This term encourages the weight distributions to maintain a degree of similarity, preventing them from becoming overly specialized in a way that harms overall performance.

The total DPR loss is a weighted sum of these two components:

$$L_{DPR} = w_{cs}L_{cs}(W_1, W_2) + w_{kl}L_{KL}(W_1, W_2), \tag{5}$$

where w_{cs} and w_{kl} are hyperparameter weights. Therefore, DPR strikes a balance: L_{cs} promotes dissimilarity in the orientation of the weight vectors for diversity, while L_{KL} prevents their distributions from becoming excessively different, thereby stabilizing the learning process and ensuring that both pathways remain effective.

3.2.3 Dual-Gaze Feature Alignment (DGFA)

To foster consistency and robustness in the learned feature space, dual-gaze feature alignment (DGFA) encourages the feature representations corresponding to the two gaze labels within a HalfMix sample (g_a, g_b) to be similar if the labels themselves are close. Let z_1 and z_2 be the normalized feature vectors (e.g., after a projection head, which is a single fully connected layer with identical input and output dimensions) from the two pathways, corresponding to predictions for g_a and g_b. The similarity between gaze labels S_{gaze} is defined using their angular distance:

$$S_{gaze}(g_a, g_b) = \exp\left(-\frac{||g_a - g_b||_2}{\tau_{gaze}}\right), \tag{6}$$

where $||g_a - g_b||_2$ is the Euclidean distance between gaze vectors, and τ_{gaze} is a temperature parameter. The similarity between features S_{feat} is their cosine similarity:

$$S_{feat}(z_1, z_2) = \frac{z_1 \cdot z_2}{||z_1||_2 \cdot ||z_2||_2}. \tag{7}$$

The DGFA loss, L_{DGFA}, minimizes the discrepancy between these similarities:

$$L_{DGFA} = \text{MSE}(S_{feat}(z_1, z_2), S_{gaze}(g_a, g_b)). \tag{8}$$

We designed DGFA based on contrastive regression losses. However, while traditional contrastive regression losses often learn relationships between different images, DGFA focuses on the relationship between the two gaze information within a single HalfMix-generated sample. This characteristic, along with the goal of directly comparing feature similarity (S_{feat}) and gaze label similarity (S_{gaze}), motivates the use of MSE. Minimizing L_{DGFA} guides the model to learn a feature space where gaze direction proximity is reflected in feature proximity, enhancing robustness to domain shifts.

The overall training objective combines these components:

$$L_{total} = L_{sup} + \alpha L_{DPR} + \beta L_{DGFA}, \tag{9}$$

where α and β are hyperparameters balancing the contributions of the regularization terms.

4 Experiments

In this section, we conduct comprehensive experiments on multiple benchmark datasets. Our evaluation encompasses both within-domain and cross-domain scenarios to assess generalization capabilities. We compare our method against baseline models and state-of-the-art approaches, followed by extensive ablation studies to analyze the contribution of each component. Additionally, we provide qualitative analysis through feature visualization to demonstrate how our method learns more robust gaze representations.

4.1 Experimental Settings

Datasets. We evaluate our method on four commonly used public gaze estimation datasets: ETH-XGaze [21] (denoted as D_E), Gaze360 [8] (denoted as D_G), MPIIFaceGaze [19] (denoted as D_M), and EyeDiap [6] (denoted as D_D). ETH-XGaze and Gaze360, which feature a large range of gaze directions, are primarily used for training and within-dataset evaluations. MPIIFaceGaze and EyeDiap are used for cross-dataset evaluation. Following standard data normalization and pre-processing steps, we apply a data rectification method [20] to normalize head poses in the EyeDiap dataset, as other datasets typically provide already rectified data or data with less head pose variation. This helps in fairer comparison by reducing performance differences solely due to head pose variations.

Implementation Details. All models were implemented using PyTorch and trained on NVIDIA A6000 GPUs. We trained all models for 10 epochs using the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a learning rate of 1×10^{-4}, and a weight decay of 1×10^{-6}. The batch size was set to 128. For our proposed method, the temperature parameter τ_{gaze} for DGFA was set to 0.07. The weights for the DPR loss components were $w_{kl} = 1$ and $w_{cs} = 0.01$. The weighting hyperparameters α and β for L_{DPR} and L_{DGFA} in the total loss (Equation 9) were both set to 1. Key hyperparameters such as τ_{gaze}, w_{kl}, w_{cs}, α, and β were determined empirically through validation experiments.

Table 1: Comparison with baseline models. Angular error (↓) in degrees. D_E and D_G are within-domain; $D_E \rightarrow D_M$, $D_E \rightarrow D_D$, $D_G \rightarrow D_M$, $D_G \rightarrow D_D$ are cross-domain. Performance improvements (red text) and degradations (blue text) relative to the baseline are shown in parentheses.

Method	D_E	$D_E \rightarrow D_M$	$D_E \rightarrow D_D$	D_G	$D_G \rightarrow D_M$	$D_G \rightarrow D_D$
ResNet18	4.98	7.34	8.13	14.55	12.81	10.46
Proposed	5.60 (-12.5%)	6.69 (8.9%)	7.97 (2.0%)	13.36 (8.2%)	8.74 (31.8%)	8.43 (19.4%)
ResNet50	4.61	7.62	8.02	14.69	14.06	12.41
Proposed	5.52 (-19.7%)	6.82 (10.5%)	6.99 (12.8%)	16.11 (-9.7%)	9.06 (35.6%)	9.14 (26.3%)

Table 2: Comparison with state-of-the-art methods. Angular error (↓) in degrees. Results for other methods are sourced from their respective publications. The 'TD' column indicates if target domain data was utilized during training (✓). Best results among methods not using target domain data (TD-free) are highlighted in **bold**.

Method	TD	$D_E \rightarrow D_M$	$D_E \rightarrow D_D$	$D_G \rightarrow D_M$	$D_G \rightarrow D_D$
GazeAdv [13]	✓	-	-	8.19	12.27
PnP-GA [11]	✓	6.00	6.17	5.74	7.04
CRGA [12]	✓	5.48	5.66	5.89	6.49
LatentGaze [10]		7.98	9.81	-	-
PureGaze [4]		7.08	7.48	9.28	9.32
AGG [1]		7.10	7.07	7.87	7.93
GazeConsistent [15]		**6.50**	7.44	7.55	**7.03**
CDG [14]		6.73	7.95	**7.03**	7.27
Ours (ResNet18)		6.69	8.31	8.74	8.43
Ours (ResNet50)		6.82	**6.99**	9.06	9.14

4.2 Comparison with Baselines

Table 1 presents a quantitative comparison of our proposed method against baseline models across various within-domain and cross-domain evaluation settings. The results clearly indicate that our full method consistently achieves substantial improvements in cross-domain gaze estimation accuracy for both ResNet18 and ResNet50 backbones. Notably, for the ResNet18 backbone, our approach reduces the angular error by up to 31.8% (in the $D_G \rightarrow D_M$ scenario), and for the ResNet50 backbone, the error reduction reaches as high as 35.6% (also for $D_G \rightarrow D_M$). While some minor performance degradations are observed in within-domain evaluations (e.g., on D_E for both backbones, and D_G for ResNet50), this is a common trade-off when optimizing for generalization. The consistent and significant gains in cross-domain settings underscore the efficacy of our proposed techniques in enhancing model robustness against domain shifts.

4.3 Comparison with state-of-the-art methods

Table 2 benchmarks our proposed method against recent state-of-the-art (SOTA) approaches in cross-domain gaze estimation. Our approach exhibits strong competitiveness, particularly among methods that do not rely on target domain data for training (TD-free). Unlike several SOTA techniques that necessitate an additional adaptation phase or access to target domain samples, our method achieves robust cross-domain generalization in a single training stage. Specifically, our ResNet50-based model sets a new SOTA for TD-free methods on the $D_E \rightarrow D_D$ task, achieving an error of 6.99 degrees. While other methods such as Gaze-

Table 3: Ablation study on the components of regularized dual-path learning. Angular error (↓) in degrees. The study incrementally adds components: Dual Path, HalfMix, DPR, and DGFA.

Method Configuration	HalfMix	Dual Path	DPR	DGFA	D_G	$D_G \rightarrow D_D$	$D_G \rightarrow D_M$
Baseline (ResNet18)					14.55	10.46	12.81
+ Dual Path		✓			14.71	13.22	11.20
+ HalfMix (Single Path)	✓				14.87	10.16	12.86
+ HalfMix + Dual Path	✓	✓			14.38	12.82	11.09
+ HalfMix + Dual Path + DPR	✓	✓	✓		14.15	9.18	10.53
+ HalfMix + Dual Path + DGFA	✓	✓		✓	13.39	8.75	9.01
+ Full	✓	✓	✓	✓	13.36	8.43	8.74

Table 4: Ablation study on different data augmentation methods. Angular error (↓) in degrees. 'ALL' refers to our full regularized dual-path learning strategy.

Method	D_G	$D_G \rightarrow D_M$	$D_G \rightarrow D_D$
Baseline (no augmentation)	14.55	12.81	10.46
Baseline + MixUp	14.74	12.36	11.13
Baseline + CutMix	14.21	13.74	13.13
Baseline + HalfMix	14.87	12.86	10.16
Baseline + HalfMix + ALL	13.36	8.74	8.43

Consistent and CDG (the TD-free part of CRGA) also demonstrate commendable results in specific settings, our proposed combination of HalfMix augmentation and regularized dual-path learning offers a compelling and effective alternative for enhancing cross-domain gaze estimation.

5 Ablation Studies

To analyze the contribution of each component of our proposed method, we conduct several ablation studies. All ablation experiments are performed using the ResNet18 backbone on the D_G dataset for within-domain evaluation, and $D_G \rightarrow D_D$ and $D_G \rightarrow D_M$ for cross-domain evaluation.

5.1 Effect of Regularized Dual-Path Learning Components

Table 3 details an ablation study investigating the contribution of each component within our regularized dual-path learning strategy. The study starts with a baseline model and incrementally adds components. The results reveal the individual and synergistic contributions of each component. Simply adding a dual-path architecture to the baseline (without HalfMix) shows mixed results. Cross-domain performance degrades on $D_G \rightarrow D_D$ (10.46 to 13.22) while slightly improving on $D_G \rightarrow D_M$ (12.81 to 11.20). Without HalfMix or DPR, the dual-path architecture leads to redundant feature learning, resulting in duplicated gradients and degraded cross-domain performance. This observation underscores the necessity of mechanisms like DPR.

Introducing HalfMix augmentation, even with a single path, generally improves cross-domain performance (e.g., $D_G \rightarrow D_D$). The combination of HalfMix and dual-path architecture further enhances this. Critically, the addition of DPR and DGFA leads to consistent improvements, particularly in cross-domain settings. For instance, adding DGFA to 'HalfMix + Dual Path' significantly reduces error on $D_G \rightarrow D_D$ (from 12.82 to 8.75) and $D_G \rightarrow D_M$

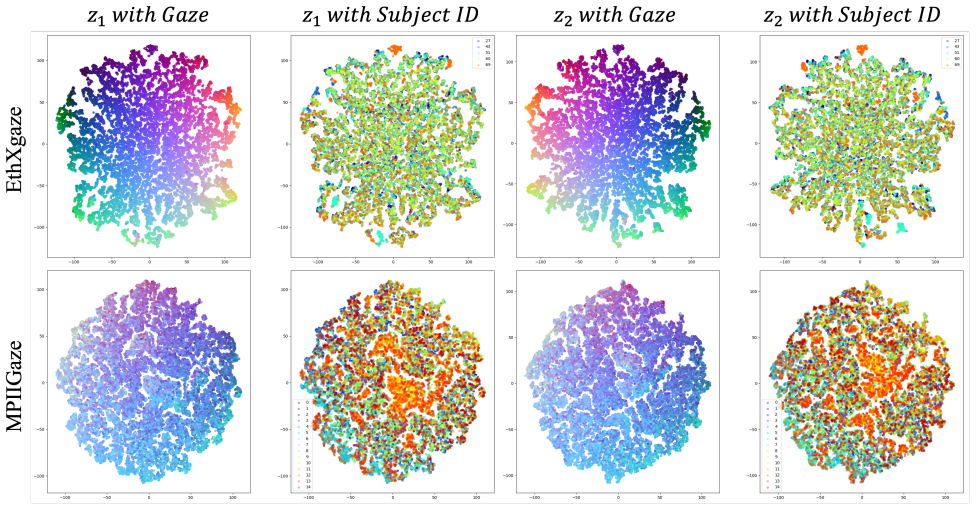

Figure 4: Feature distribution from our full method, using z_1 and z_2. Features show improved alignment with gaze direction and reduced subject-specific clustering.

(from 11.09 to 9.01). The full model, incorporating all components, achieves the best overall cross-domain performance, highlighting the importance of each element in our regularized dual-path learning strategy.

5.2 Effect of Different Data Augmentation Methods

Table 4 illustrates the impact of different augmentation strategies. While standard MixUp and CutMix show varied, and sometimes detrimental, effects on cross-domain performance compared to the baseline, HalfMix alone (Baseline + HalfMix) demonstrates a notable improvement, especially on the $D_G \rightarrow D_D$ task (10.46 to 10.16). More significantly, when HalfMix is integrated with our complete regularized dual-path learning strategy (Baseline + HalfMix + ALL, our full method), we observe substantial performance gains across all evaluated scenarios, particularly in cross-domain settings ($D_G \rightarrow D_M$: 12.81 to 8.74; $D_G \rightarrow D_D$: 10.46 to 8.43). This underscores that HalfMix, when coupled with its tailored learning strategy, is considerably more effective for cross-domain gaze estimation than naive applications of other mix-based methods.

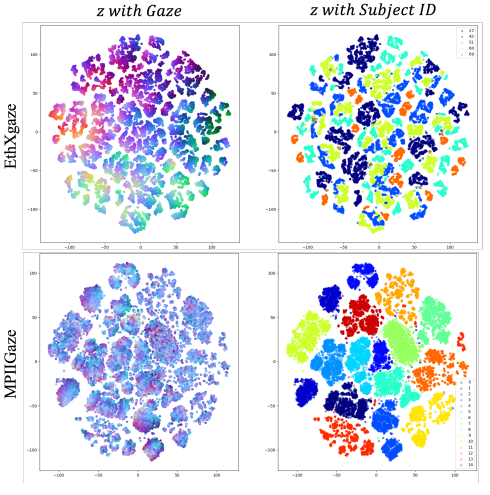

Figure 3: Feature distribution from baseline model, using f_{common}. Features tend to cluster by subject ID rather than gaze direction.

5.3 Qualitative Analysis

To further understand how our proposed method contributes to cross-domain generalization, we visualize the learned feature distributions using t-SNE. Figure 3 shows the t-SNE visualization of the common features f_{common} learned by a baseline model (e.g., ResNet18 without our proposed HalfMix and regularized dual-path learning). It can be observed that these features tend to cluster based on subject identity rather than gaze direction, which can hinder generalization to unseen subjects. In contrast, Figure 4 illustrates the t-SNE visualization of features learned by our full method, using the feature vectors z_1 and z_2 from the two pathways. The features exhibit a much clearer alignment with gaze direction, forming a more structured manifold that is less sensitive to subject-specific characteristics. This improved feature disentanglement and organization are crucial for enhancing cross-domain gaze estimation performance, as the model learns representations that are more inherently tied to the gaze itself rather than spurious, domain-specific cues.

6 Limitations

While our method demonstrates strong cross-domain performance, several limitations should be acknowledged. HalfMix assumes both eye regions are visible, which may limit its effectiveness for extreme head poses where self-occlusion occurs. The observed minor within-domain performance degradation suggests room for adaptive mixing strategies that could balance generalization and in-domain accuracy. Additionally, the dual-path architecture introduces computational overhead during training compared to single-path models. Future work could explore adaptive splitting strategies for challenging head poses and develop more efficient architectures while maintaining the generalization benefits.

7 Conclusion

In this paper, we presented HalfMix augmentation and a regularized dual-path learning strategy to tackle the critical challenge of cross-domain gaze estimation. HalfMix effectively generates diverse training samples by combining two images while preserving crucial eye regions, mitigating issues common in other mix-based augmentations. The regularized dual-path learning strategy, with its dual-path architecture, diversity-promoting regularization, and dual-gaze feature alignment, enables the model to learn robust and generalizable features from these augmented samples. Our extensive experiments demonstrated that the proposed method significantly outperforms existing approaches in various cross-domain scenarios, achieving state-of-the-art results among target domain data-free methods on several benchmarks. Qualitative analysis further confirmed that our method learns a feature space better aligned with gaze direction and less sensitive to subject-specific characteristics.

Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2025-02283048, Developing the Next-Generation General AI with Reliability, Ethics, and Adaptability). Furthermore, this work was also supported in part by the Core Research Institute Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2021-NR060127).

References

[1] 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*, pages 1409–1418, 2024.

[2] Alisa Burova, John Mäkelä, Jaakko Hakulinen, Tuuli Keskinen, Hanna Heinonen, Sanni Siltanen, and Markku Turunen. Utilizing vr and gaze tracking to develop ar solutions for industrial maintenance. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020.

[3] Nora Castner, Thomas C Kuebler, Katharina Scheiter, Juliane Richter, Thérése Eder, Fabian Hüttig, Constanze Keutel, and Enkelejda Kasneci. Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing. In *ACM symposium on eye tracking research and applications*, pages 1–10, 2020.

[4] Yihua Cheng, Yiwei Bao, and Feng Lu. PureGaze: Purifying Gaze Feature for Generalizable Gaze Estimation. https://arxiv.org/abs/2103.13173v2, March 2021.

[5] 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, February 2024. ISSN 1939-3539. doi: 10.1109/TPAMI.2024.3393571.

[6] 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*, pages 255–258, 2014.

[7] Seong-Hyeon Hwang and Steven Euijong Whang. RegMix: Data Mixing Augmentation for Regression, August 2022.

[8] 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*, pages 6912–6921, 2019.

[9] 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.

[10] 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 (ACCV)*, pages 3379–3395, December 2022.

[11] 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*, pages 3835–3844, 2021.

[12] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Aaron Courville, David Lopez-Paz, and Yoshua Bengio. Manifold Mixup: Better Representations by Interpolating Hidden States, May 2019.

[13] 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*, pages 11907–11916, 2019.

[14] 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*, pages 19376–19385, 2022.

[15] Mingjie Xu, Haofei Wang, and Feng Lu. Learning a Generalized Gaze Estimator from Gaze-Consistent Feature. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):3027–3035, June 2023. ISSN 2374-3468. doi: 10.1609/aaai.v37i3.25406.

[16] Huaxiu Yao, Yiping Wang, Linjun Zhang, James Zou, and Chelsea Finn. C-Mixup: Improving Generalization in Regression, October 2022.

[17] 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 (ICCV)*, October 2019.

[18] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization, April 2018.

[19] 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*, pages 51–60, 2017.

[20] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pages 1–9, 2018.

[21] 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 Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 365–381, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58558-7. doi: 10.1007/978-3-030-58558-7_22.