LEARNING TO GAZE BEYOND THE FACE: UNSUPERVISED TRANSFER FROM CROPPED-FACE EXPERTS TO FULL-FRAME DRIVING SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale gaze estimation datasets are predominantly built from cropped face images, making their models difficult to apply directly to full-frame driving scenes. We address this cropped-face-to-full-frame gap in the absence of target-domain gaze annotations. Our framework retains a pretrained cropped-face gaze expert as a source of privileged gaze knowledge and trains a full-frame model without explicit feature alignment. Instead, multiple complementary objectives transfer prediction-level, relational, and gaze-sensitive knowledge while enforcing consistency and an anti-collapse constraint to preserve the intrinsic variation of gaze predictions. During inference, the privileged expert is removed and gaze is predicted directly from raw full-frame images. Experiments on in-vehicle gaze estimation show that the proposed approach substantially outperforms direct expert transfer and other knowledge transfer baselines, demonstrating effective annotation-free transfer from cropped faces to full-frame driving scenes.

1 INTRODUCTION

Gaze estimation plays an important role in understanding human visual attention and has been widely explored in applications such as human-computer interaction Aftab (2019); Murali et al. (2022), saliency detection Wang et al. (2019), and driver monitoring Hu et al. (2025a). In driving scenarios, estimating the driver's gaze from in-vehicle cameras provides a direct cue for monitoring visual attention and detecting potentially risky behaviors Hu et al. (2025b). Despite substantial progress in recent years, practical gaze estimation systems still face a fundamental gap between existing benchmarks and real-world deployment Huang et al. (2025).

Most existing gaze estimation methods are developed using large-scale datasets collected from cropped or normalized face images Zhang et al. (2015); Kellnhofer et al. (2019); Zhang et al. (2020). This setting enables the model to focus on gaze-related facial appearance, but it also introduces an additional dependency on face localization and cropping during inference. In contrast, an in-vehicle camera naturally captures a raw full-frame image containing the face together with a large amount of background and scene information. Applying a crop-based gaze estimator to such images therefore requires an additional face detection or landmark localization stage, which introduces extra complexity and may lead to performance degradation under challenging driving conditions. Although end-to-end frame-to-gaze estimation has demonstrated the feasibility of directly predicting gaze from raw frames Balim et al. (2023), training such models typically requires corresponding supervision in the target setting.

A more practical scenario is to reuse the gaze knowledge already learned from existing large-scale cropped-face datasets and transfer it to an unlabeled raw-frame driving domain. However, this problem is non-trivial because the source and target observations are fundamentally different. A cropped face provides a compact and gaze-relevant representation, whereas a raw frame contains both the face and substantial irrelevant visual content. We therefore regard the cropped face as a privileged facial observation, namely, an additional face-centered input available during training but not required at inference. Directly using predictions from a crop-based expert as pseudo labels is also unreliable, since the cropped faces extracted from the target domain may differ considerably from the source training distribution. Consequently, the key challenge is how to transfer the gaze

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.

knowledge provided by this privileged observation to a raw-frame model without requiring target gaze annotations.

To address this challenge, we propose a privileged observation-based knowledge transfer framework for crop-free gaze estimation. A gaze expert pretrained on cropped-face images is used only during training, while a full-frame student is optimized without target gaze labels. The cropped face is treated as a privileged observation that provides complementary gaze-related information unavailable to the student at inference time. Rather than relying solely on direct pseudo-label supervision, our framework transfers complementary knowledge from the privileged observation to the raw-frame student and regularizes its prediction behavior to improve the reliability of full-frame gaze estimation.

Extensive experiments on the LBW driving dataset demonstrate that our approach substantially outperforms existing crop-to-frame knowledge transfer baselines under standard, high-illumination, and nighttime conditions. Moreover, the proposed framework remains effective across different gaze expert sources, unlabeled-data proportions, and privileged observation settings, demonstrating its robustness and general applicability to crop-free gaze estimation. The main contributions of this work are summarized as follows:

- We formulate crop-free gaze estimation as knowledge transfer from privileged facial observations to raw full-frame inputs without target-domain gaze annotations.
- We propose a privileged observation-based framework that transfers complementary gaze knowledge while regularizing full-frame predictions without explicit feature alignment.
- Extensive experiments demonstrate consistent improvements over existing transfer baselines across diverse driving conditions and experimental settings.

2 Related Work

2.1 Gaze Estimation

Appearance-based gaze estimation has been extensively studied using large-scale in-the-wild datasets Qu et al. (2026); Vuillecard & Odobez (2025). MPIIGaze established the importance of diverse real-world gaze data Zhang et al. (2015), while Gaze360 further introduced unconstrained variations in subjects, head poses, and viewing distances Kellnhofer et al. (2019). ETH-XGaze substantially increased the scale and coverage of gaze and head-pose variations with more than one million high-resolution images Zhang et al. (2020). Recent studies have further explored domain generalization and adaptation, including outlier-guided adaptation Liu et al. (2021), contrastive regression Wang et al. (2022), rotation consistency Bao et al. (2022), uncertainty reduction Cai et al. (2023), and improved regression formulations Bao & Lu (2024). Despite these advances, most existing gaze estimators operate on cropped or normalized facial observations rather than raw frames.

A fundamental limitation of this setting is that inference typically requires an additional face localization and cropping procedure before gaze estimation. EFE Balim et al. (2023) explored end-to-end frame-to-gaze estimation by directly predicting 3D gaze information from raw camera frames, demonstrating the feasibility of avoiding explicit face cropping. However, EFE focuses on directly learning a frame-to-gaze mapping from datasets containing the corresponding training supervision, rather than transferring gaze knowledge from existing large-scale cropped-face datasets to an unlabeled raw-frame target domain. In our setting, the challenge is therefore to exploit existing face-based gaze knowledge while removing the crop dependency at inference.

2.2 Knowledge Transfer

Knowledge distillation (KD) was originally introduced to transfer knowledge from a strong teacher to a student through softened output predictions Hinton et al. (2015). Subsequent studies extended this paradigm to intermediate representations Romero et al., relational structures Park et al. (2019), and other transferable cues such as feature similarity and attention Yim et al. (2017); Zagoruyko & Komodakis (2016); Liu et al. (2019). These studies established knowledge transfer as a general paradigm for extracting richer supervision from pretrained models.

Recent research has increasingly considered heterogeneous teacher–student settings, where the two models differ in architecture, input characteristics, or representation space. Several works have explored transfer across heterogeneous representations or input spaces Passalis et al. (2020); Huo et al. (2024), while adaptive strategies have been proposed to mitigate teacher–student discrepancies Zhao et al. (2020). PAT Lin et al. (2025), G^{2}D Rakib & Bagavathi (2025), MonoTAKD Liu et al. (2025), and SKD Jeong et al. (2026) further investigate structured knowledge transfer under heterogeneous settings.

However, these methods were not designed for cropped-face-to-full-frame gaze transfer without target-domain gaze annotations. For fair comparison, we adapt their core transfer mechanisms to our setting, using the frozen cropped-face expert as the teacher and the raw full-frame model as the student, without target gaze supervision. In our setting, the cropped face is a privileged observation available only during training and may provide noisy target-domain predictions. Therefore, our framework transfers gaze-related knowledge from this privileged observation without directly enforcing conventional teacher–student alignment.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERALL FRAMEWORK

Existing large-scale gaze estimation datasets are predominantly constructed from cropped face images, which suppress irrelevant scene content and allow gaze models to focus on facial appearance and head pose. In practical driver monitoring, however, in-vehicle cameras capture complete scenes containing the driver and surrounding environment. Collecting large-scale gaze annotations under this setting is costly. We therefore study the transfer of gaze knowledge from labeled cropped-face data to unlabeled raw full-frame driving scenes.

Let $\mathcal{X}^c$ and $\mathcal{X}^r$ denote the cropped-face and raw full-frame image domains, respectively. A cropped face is denoted by $x^c \in \mathcal{X}^c$, while $x^r \in \mathcal{X}^r$ denotes a complete driving scene. We first train a gaze expert $T(\cdot)$ on labeled cropped-face data and freeze it during transfer. Given x^r, a facial landmark based cropping operator $C(\cdot)$ produces $x^c = C(x^r)$. This crop is used only as a privileged observation during training.

Directly using $T(C(x^r))$ as a pseudo label is unreliable because target-domain crops may differ from the expert's training distribution due to illumination, camera characteristics, image quality, subject appearance, and localization errors. We therefore regard the target-domain crop as a noisy privileged observation containing both transferable gaze knowledge and domain-specific uncertainty. The goal is to transfer stable gaze-related information rather than reconstructing the cropped-face representation.

Accordingly, we train a full-frame model $S_\theta(\cdot)$ together with the frozen expert. For x^r, the student produces $g^S = S_\theta(x^r)$, while the expert produces $g^T = T(C(x^r))$. Multiple appearance perturbations are used to estimate the reliability of privileged predictions, followed by complementary transfer at the prediction, gaze-space relational, and gaze-sensitive levels, together with full-frame consistency and anti-collapse regularization. During inference, the privileged branch is removed and only $S_\theta(\cdot)$ is retained.

3.2 MODELING NOISY PRIVILEGED KNOWLEDGE

For each target-domain face crop, we construct multiple lightweight appearance perturbations. Let $\{\mathcal{A}_k\}_{k=1}^K$ denote K stochastic transformations, such as Gaussian noise, salt-and-pepper noise, color perturbation, or brightness variation. For the i-th sample, $x_i^c = C(x_i^r)$ and $x_{i,k}^c = \mathcal{A}_k(x_i^c)$. The frozen expert produces $g_{i,k}^T = T(x_{i,k}^c)$. The mean prediction represents the central expert response, while its variance reflects sensitivity to appearance changes.

Let $\bar{g}_i^T$ denote the mean of the K expert predictions. For the d-th gaze dimension, the prediction variance is $\mathrm{Var}_k(g_{i,k,d}^T)$. We convert this uncertainty into a confidence coefficient c_i:

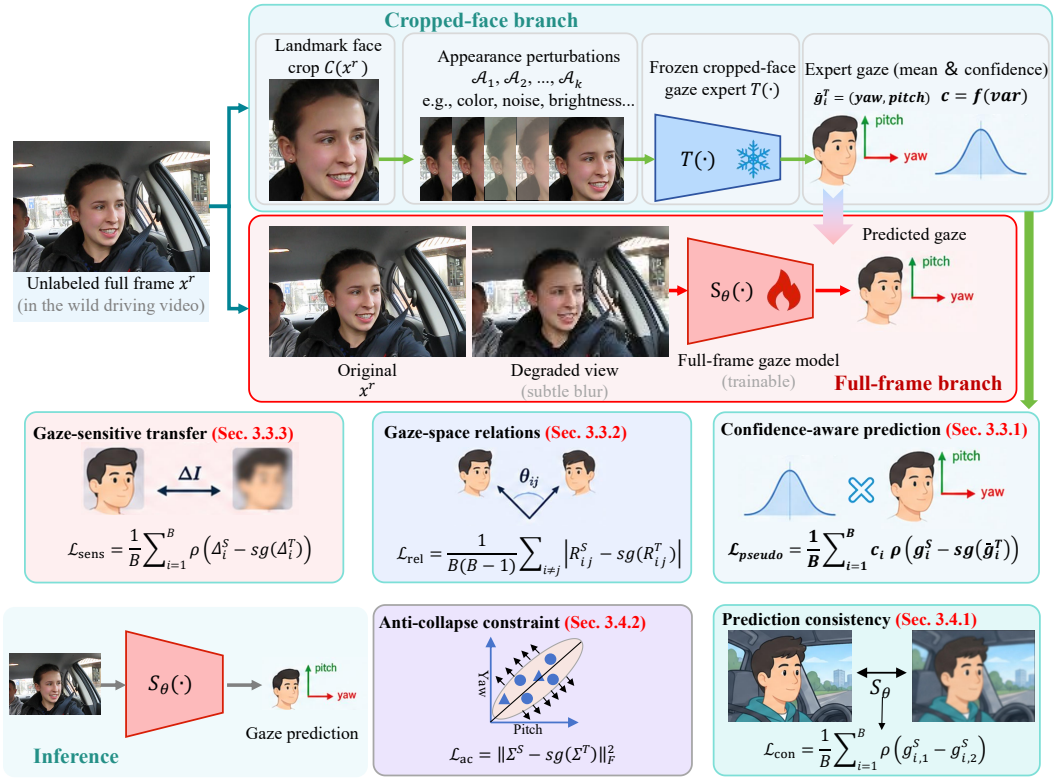


Figure 1: Overview of the proposed privileged observation-based knowledge transfer framework. A frozen cropped-face gaze expert provides noisy privileged knowledge for training a full-frame gaze model through multi-level gaze transfer and full-frame self-supervised regularization. The privileged branch is removed during inference.

$$c_i = \exp\left(-\frac{1}{D}\sum_{d=1}^{D}\mathrm{Var}_k\left(g_{i,k,d}^T\right)\right), \tag{1}$$

where D denotes the dimensionality of the gaze prediction, corresponding to yaw and pitch. A larger c_i indicates a more stable expert response.

3.3 Multi-level Gaze Knowledge Transfer

3.3.1 Confidence-aware Prediction Transfer

The expert prediction provides direct gaze guidance, but target-domain predictions may contain noise. We therefore weight the transfer according to the stability of the privileged prediction.

Given a mini-batch $\mathcal{B} = \{x_i^r\}_{i=1}^B$, the full-frame prediction is $g_i^S = S_\theta(x_i^r)$, while $\bar{g}_i^T$ is the mean expert prediction from the perturbed face observations. Let $\rho(\cdot)$ denote the Smooth-L_1 robust regression function. The confidence-aware pseudo objective is

$$\mathcal{L}_{\text{pseudo}} = \frac{1}{B}\sum_{i=1}^{B} c_i\, \rho\left(g_i^S - \mathrm{sg}\left(\bar{g}_i^T\right)\right), \tag{2}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator.

3.3.2 GAZE-SPACE RELATIONAL TRANSFER

Point-wise prediction matching may inherit systematic biases from the expert. We therefore preserve the relative geometry among gaze predictions.

For a two-dimensional gaze prediction $g = (y, p)$, where y and p denote yaw and pitch, respectively, we map it to a three-dimensional unit gaze vector $\mathbf{v} = \phi(y, p)$. The mapping used in our implementation is $\mathbf{v} = (-\cos p \sin y, -\sin p, -\cos p \cos y)$ followed by unit normalization. For two samples i and j, the gaze relation is defined as $\mathcal{R}(\mathbf{v}_i, \mathbf{v}_j) = \arccos(\mathrm{clip}(\mathbf{v}_i^\top \mathbf{v}_j, -1+\epsilon, 1-\epsilon))$, where ϵ avoids numerical instability.

The expert and full-frame predictions form relational matrices $\mathbf{R}^T$ and $\mathbf{R}^S$, respectively. Their discrepancy is minimized by

$$\mathcal{L}_{\mathrm{rel}} = \frac{1}{B(B-1)} \sum_{i \neq j} \left| R_{ij}^S - \mathrm{sg}(R_{ij}^T) \right|. \tag{3}$$

This transfers the relative organization of gaze predictions rather than intermediate visual representations.

3.3.3 GAZE-SENSITIVE RESPONSE TRANSFER

The eye occupies only a small region of the full frame, while fine facial and ocular details provide important gaze cues. We therefore transfer the prediction response induced by controlled visual degradation.

Let $\mathcal{P}(\cdot)$ denote a detail degradation operator, instantiated by 3×3 average pooling in our implementation. The original and degraded inputs produce g_i^S and $\tilde{g}_i^S$ for the full-frame model, and $\bar{g}_i^T$ and $\tilde{g}_i^T$ for the expert. Their prediction responses are $\Delta_i^S = g_i^S - \tilde{g}_i^S$ and $\Delta_i^T = \bar{g}_i^T - \tilde{g}_i^T$. The resulting sensitivity objective is

$$\mathcal{L}_{\mathrm{sens}} = \frac{1}{B} \sum_{i=1}^{B} \rho\left(\Delta_i^S - \mathrm{sg}\left(\Delta_i^T\right)\right). \tag{4}$$

3.4 FULL-FRAME SELF-SUPERVISED REGULARIZATION

3.4.1 PREDICTION CONSISTENCY

The raw full-frame images provide additional self-supervised information. Two independently perturbed versions $x_{i,1}^r$ and $x_{i,2}^r$ are generated for each image, producing $g_{i,1}^S = S_\theta(x_{i,1}^r)$ and $g_{i,2}^S = S_\theta(x_{i,2}^r)$. The consistency objective is

$$\mathcal{L}_{\mathrm{con}} = \frac{1}{B} \sum_{i=1}^{B} \rho\left(g_{i,1}^S - g_{i,2}^S\right). \tag{5}$$

In our implementation, Gaussian noise is used to generate the two perturbed full-frame inputs.

3.4.2 ANTI-COLLAPSE CONSTRAINT

Consistency regularization may admit a degenerate solution with nearly identical predictions for all inputs. We therefore preserve the non-degenerate variation of gaze predictions within each batch.

For the full-frame predictions $\{g_i^S\}_{i=1}^B$, let μ^S and Σ^S denote their mean and covariance matrix. Similarly, μ^T and Σ^T denote the corresponding expert statistics. The student covariance is

$\Sigma^S = (B-1)^{-1} \sum_{i=1}^B (g_i^S - \mu^S)(g_i^S - \mu^S)^\top$. We constrain the two covariance structures through

$$\mathcal{L}_{\mathrm{ac}} = \left\| \Sigma^S - \mathrm{sg}\left(\Sigma^T\right) \right\|_F^2, \tag{6}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

3.5 Overall Optimization

The five objectives provide complementary supervision: $\mathcal{L}_{\text{pseudo}}$ provides sample-level gaze guidance, $\mathcal{L}_{\text{rel}}$ preserves gaze-space geometry, $\mathcal{L}_{\text{sens}}$ transfers gaze-sensitive responses, while $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{ac}}$ regularize full-frame predictions.

The final optimization objective is

$$\mathcal{L} = \lambda_{\text{pseudo}}\mathcal{L}_{\text{pseudo}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}} + \lambda_{\text{sens}}\mathcal{L}_{\text{sens}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{ac}}\mathcal{L}_{\text{ac}}, \tag{7}$$

where λ_{pseudo}, λ_{rel}, λ_{sens}, λ_{con}, and λ_{ac} denote the weights of the five objectives. In our implementation, $\lambda_{\text{pseudo}} = 0.5$, $\lambda_{\text{rel}} = 0.5$, $\lambda_{\text{sens}} = 0.5$, and $\lambda_{\text{con}} = 0.1$, while the anti-collapse coefficient is controlled separately by the training configuration. No target-domain gaze annotations are used during full-frame training.

During inference, the privileged face crop and frozen expert are removed. Given a raw full-frame image x^r, the final model directly predicts gaze as $\hat{g} = S_\theta(x^r)$.

4 Experiments

4.1 Experimental Setup

Datasets and Protocol. The gaze expert is pretrained on ETH-XGaze Zhang et al. (2020) and kept frozen during transfer. Unless otherwise specified, ETH-XGaze is used as the source domain, while Gaze360 Kellnhofer et al. (2019) is used only in designated experiments. LBW Kasahara et al. (2022) serves as the unlabeled target driving domain. During training, facial landmarks are used only to obtain privileged face crops, and the full-frame model takes both the raw image and privileged crop as input. The complete LBW test set is held out and unseen during training. For evaluation, LBW-L contains the top 10% samples by mean image brightness and represents overexposed scenes, while LBW-N consists of manually selected nighttime scenes. At inference, only the raw full-frame image is provided. We report mean angular error computed from normalized 3D gaze vectors converted from yaw and pitch, with lower values indicating better performance.

Implementation Details. The proposed method is implemented by using Pytorch, and trained for 50 epochs on two V100 GPUs. A batch-size of 256 is set to train, and the learning rate is set as 0.00001. The learning rate is adjusted every 25 epochs with an adjustment coefficient is 0.1. The gaze expert remains frozen throughout training.

4.2 Main Comparison

Table 1 compares our method with representative approaches under standard, high-illumination, and nighttime driving conditions. All methods use ResNet18 for a fair comparison. The Source-only Expert suffers from a substantial cropped-face-to-full-frame domain gap, with the angular error increasing from 23.43° on LBW to 29.18° on LBW-N. Teacher Pseudo Supervision reduces the error to 17.66°, showing that the privileged crop provides useful but noisy gaze knowledge.

Existing knowledge transfer methods further improve the performance, with SKD achieving the strongest baseline at 14.76° on LBW. Our method achieves the best performance across all three conditions, reaching 11.25°, 13.58°, and 14.13° on LBW, LBW-L, and LBW-N, respectively. Compared with SKD, this yields reductions of 3.51°, 3.25°, and 4.79°. The consistent gains across high-illumination and nighttime conditions also suggest improved robustness to full-frame appearance variations.

The $< 10^\circ$ accuracy further confirms this improvement. Our method achieves 50.62% on LBW, substantially exceeding SKD at 27.05%, and maintains clear advantages on LBW-L and LBW-N with 39.76% and 36.84%, respectively. Although the fully supervised model remains superior, the consistent gains in both angular error and low-error prediction ratio demonstrate the effectiveness of the proposed privileged knowledge transfer framework.

Table 1: Comparison with representative knowledge transfer methods under different driving conditions. All methods use ResNet18 as the backbone.

Method	LBW		LBW-L		LBW-N	
	Angle ↓	$< 10^\circ$ ↑	Angle ↓	$< 10^\circ$ ↑	Angle ↓	$< 10^\circ$ ↑
Supervised	9.12	69.27%	11.06	61.48%	12.27	54.36%
Source-only Expert	23.43	12.84%	26.15	10.37%	29.18	9.91%
Teacher Pseudo Supervision	17.66	18.59%	20.41	13.74%	22.93	10.21%
PAT Lin et al. (2025)	15.84	26.39%	18.27	18.82%	20.38	14.73%
G^{2}D Rakib & Bagavathi (2025)	16.21	22.60%	18.86	16.91%	21.17	13.11%
MonoTAKD Liu et al. (2025)	15.47	22.47%	17.91	17.22%	20.11	13.76%
SKD Jeong et al. (2026)	14.76	27.05%	16.83	20.94%	18.92	16.42%
Ours	**11.25**	**50.62%**	**13.58**	**39.76%**	**14.13**	**36.84%**

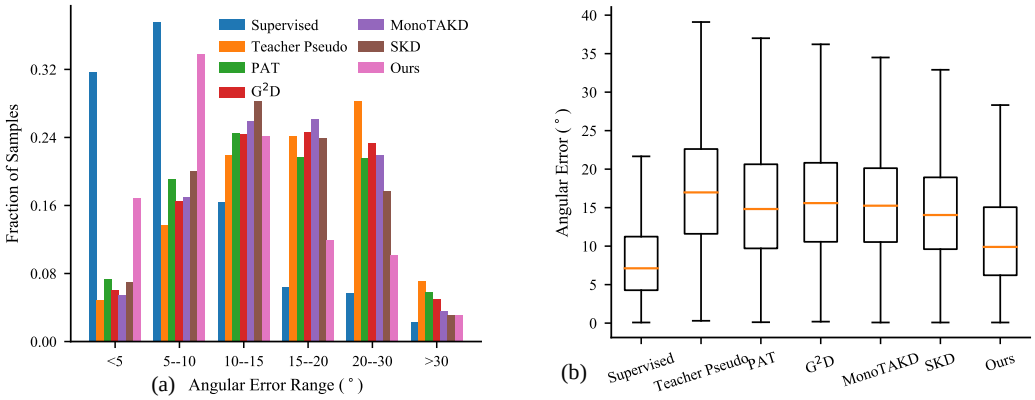


Figure 2: Distribution of angular errors on the LBW test set. (a) Proportion of samples within different angular-error ranges. (b) Boxplots of angular errors, where the box indicates the interquartile range, the center line denotes the median, and the whiskers represent the non-outlier error range.

4.3 Error Distribution Analysis

Figure 2(a) shows the proportion of samples within different angular-error ranges. Our method achieves the second-highest proportion of low-error predictions in the 0°–10° range, substantially exceeding the knowledge transfer baselines, while remaining below the supervised model. The gap becomes more evident in the stringent $< 5^\circ$ range, indicating that the transferred knowledge still cannot fully recover the accuracy obtained with direct target-domain supervision.

Figure 2(b) further examines the overall error distribution. Compared with the non-supervised transfer baselines, our method achieves a lower median error together with a more compact interquartile range and a shorter upper whisker, indicating fewer relatively large errors. Although the supervised model remains superior overall, the consistent improvement in both low-error sample proportion and error dispersion demonstrates that our method improves not only the average performance but also the distribution of prediction errors.

4.4 Error Distribution Analysis

Figure 3(a) compares the density distributions of angular errors on the LBW test set. Our method exhibits a clear shift toward the low-error region compared with the knowledge transfer baselines, with only the supervised model showing a more favorable distribution. In contrast, Teacher Pseudo Supervision retains more probability mass in the high-error region. This distributional trend is consistent with the average angular errors in Table 1, indicating that the improvement of our method is reflected across a broader range of samples.

Figure 3(b) further compares the full-frame and corresponding cropped-face predictions through their error difference. Our method shows a stronger shift toward negative values than the other

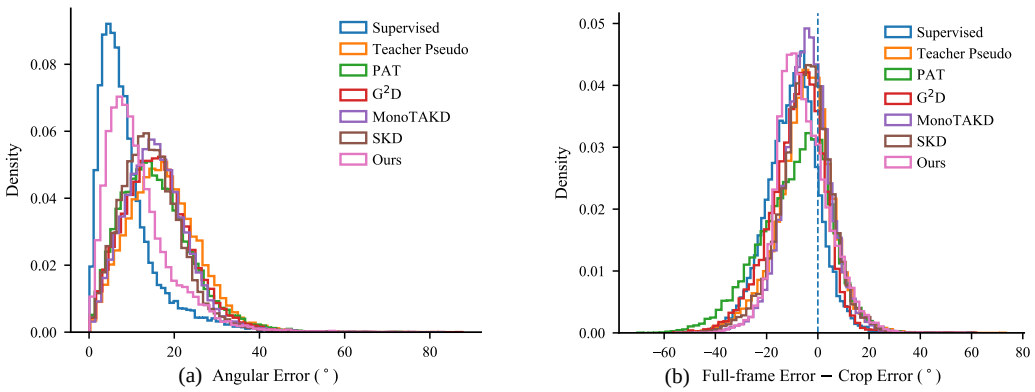


Figure 3: Error distribution analysis on the LBW test set. (a) Density distribution of angular errors. A leftward shift indicates a larger concentration of low-error predictions. (b) Distribution of the error difference between full-frame and cropped-face predictions, where $\Delta E = E_{\text{full}} - E_{\text{crop}}$. Negative values indicate lower error for the full-frame prediction.

Table 2: Comparison of different gaze knowledge transfer strategies across different backbones.

Backbone	Method	LBW	LBW-L	LBW-N
VGG16	Full-Frame Model (Supervised)	9.37	12.11	12.58
	Frozen Face Expert	25.07	27.84	30.26
	Teacher Pseudo Supervision	18.93	21.46	23.87
	Feature-level Alignment	17.41	19.86	22.08
	Ours	12.18	14.36	15.02
ResNet18	Full-Frame Model (Supervised)	9.12	11.06	12.27
	Frozen Face Expert	23.43	26.15	29.18
	Teacher Pseudo Supervision	17.66	20.41	22.93
	Feature-level Alignment	16.37	18.74	21.06
	Ours	11.25	13.58	13.73
GazeTR	Full-Frame Model (Supervised)	8.76	10.72	11.14
	Frozen Face Expert	22.71	25.63	28.54
	Teacher Pseudo Supervision	17.21	19.93	22.41
	Feature-level Alignment	15.98	18.36	20.71
	Ours	10.92	13.21	13.86

methods, indicating that full-frame predictions achieve lower errors than the corresponding crop predictions for a larger portion of samples. This suggests that the proposed transfer strategy does not simply reproduce the privileged crop prediction, but enables the full-frame model to learn a more effective gaze representation from the transferred knowledge.

4.5 Comparison with Crop-to-Frame Transfer Baselines

Table 2 evaluates different strategies for transferring cropped-face gaze knowledge to raw full-frame images. Across all backbones, the Frozen Face Expert performs substantially worse than the fully supervised full-frame model, while Teacher Pseudo Supervision provides a clear improvement, confirming the transferability of the pretrained gaze knowledge. Feature-level alignment further improves the results but remains inferior to our method, suggesting that feature matching alone cannot fully address the crop-to-frame discrepancy.

Our method consistently achieves the best label-free transfer performance across different backbones and driving conditions. With ResNet18, the angular error decreases from $17.66°$ with Teacher Pseudo Supervision and $16.37°$ with feature-level alignment to $11.25°$, approaching the fully supervised reference of $9.12°$. Similar trends are observed on LBW-L and LBW-N. While VGG16 performs relatively worse, GazeTR and ResNet18 show comparable results, indicating that the proposed improvement is not strongly dependent on the backbone architecture.

Table 3: Ablation study of the proposed training objectives on LBW.

Configuration	Angle (°) ↓
Pseudo Only	17.42
Relation Only	18.61
Sensitivity Only	18.03
Pseudo + Relation	13.74
Pseudo + Sensitivity	16.03
Pseudo + Consistency	16.61
Pseudo + Anti-collapse	16.38
Pseudo + Relation + Sensitivity	12.61
Pseudo + Relation + Consistency + Anti-collapse	11.83
Full Framework	**11.25**

Table 4: Transferability of gaze experts pretrained on different source datasets.

Teacher Source	Teacher	Pseudo Supervision	Ours
ETH-XGaze	16.92	17.66	**11.25**
Gaze360	17.34	18.03	**12.47**

4.6 Ablation Study

The results in Table 3 show that pseudo supervision provides a strong absolute gaze reference, while relation and sensitivity alone are less effective. Notably, adding relation to pseudo supervision reduces the error from $17.42°$ to $13.74°$, demonstrating the importance of preserving gaze-space structure beyond point-wise supervision. Sensitivity provides a complementary gain, further reducing the error to $12.61°$ when combined with relation. Consistency and anti-collapse provide additional regularization, with their combination improving the error to $11.83°$. Finally, jointly optimizing all five objectives achieves the best performance of $11.25°$, confirming their complementary roles in the proposed transfer framework.

4.7 Transfer from Different Gaze Expert Sources

To examine whether the proposed framework depends on a specific source gaze dataset, we pretrain the face expert using two different large-scale gaze datasets, ETH-XGaze and Gaze360, and transfer the resulting knowledge to the same LBW target domain. All other components and training settings are kept unchanged. Table 4 shows that the proposed method remains effective when the source gaze expert is pretrained on different large-scale datasets. Although the Gaze360-based expert yields slightly weaker target-domain predictions than the ETH-XGaze-based expert, the corresponding full-frame model still achieves a comparable result. The relatively small performance variation suggests that the proposed framework is not strongly tied to a particular source dataset, but can transfer gaze knowledge from different pretrained face experts.

5 Conclusion

In this work, we proposed an unsupervised cropped-face-to-full-frame gaze transfer framework for in-vehicle gaze estimation. By treating the cropped face as a privileged observation, the proposed method transfers complementary gaze knowledge through pseudo supervision, relation modeling, sensitivity transfer, consistency regularization, and anti-collapse constraint, without requiring target-domain gaze annotations during training. Extensive experiments on LBW under normal, high-illumination, and nighttime conditions demonstrate consistent improvements over direct expert transfer and existing knowledge transfer baselines. The results show that the proposed framework enables a full-frame model to learn effective gaze representations from privileged cropped-face knowledge while avoiding explicit feature-level alignment.

AI USE STATEMENT

In this work, we used generative AI tools to assist with software code development and to improve the grammar, wording, and linguistic clarity of the manuscript. The code developed with AI assistance was manually reviewed, checked, and repeatedly verified by the authors. As non-native English speakers, we also used large language models to identify and correct grammatical errors and inappropriate word usage. All methodological decisions, experimental design, result interpretation, and scientific claims were determined and verified by the authors. We did not use generative AI tools to generate scientific figures or images; all figures in the manuscript, including the methodology framework, were manually created as vector graphics by the authors. No synthetic data, AI-generated experimental results, or AI-generated research contributions were used in this work. We reviewed all AI-assisted content and take full responsibility for the final content of this work, including all text, code, claims, results, and figures.

REFERENCES

Abdul Rafey Aftab. Multimodal driver interaction with gesture, gaze and speech. In *2019 International Conference on Multimodal Interaction*, pp. 487–492, 2019.

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

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

Yiwei Bao, Yunfei Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with rotation consistency. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4197–4206. IEEE, 2022.

Xin Cai, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Source-free adaptive gaze estimation by uncertainty reduction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22035–22045. IEEE, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Daosong Hu, Mingyue Cui, and Kai Huang. Unmode: Uncertainty modeling for driver gaze estimation via feature disentanglement. *IEEE Transactions on Intelligent Transportation Systems*, 26 (7):10612–10622, 2025a.

Daosong Hu, Xi Li, Mingyue Cui, and Kai Huang. Lnet: Lightweight network for driver attention estimation via scene and gaze consistency. *IEEE Transactions on Image Processing*, 35:27–41, 2025b.

Yexin Huang, Yongbin Lin, Lishengsa Yue, Zhihong Yao, and Jie Wang. From gaze to movement: predicting visual attention for autonomous driving human-machine interaction based on programmatic imitation learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26146–26155. IEEE, 2025.

Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C 2 kd: Bridging the modality gap for cross-modal knowledge distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16006–16015. IEEE, 2024.

Hyeon-Jin Jeong, Han-Jin Lee, and Seok-Hwan Choi. Streamlined knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26646–26655, 2026.

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.

Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6911–6920. IEEE, 2019.

Jhe-Hao Lin, Yi Yao, Chan-Feng Hsu, Hong-Xia Xie, Hong-Han Shuai, and Wen-Huang Cheng. Perspective-aware teaching: Adapting knowledge for heterogeneous distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4178–4187. IEEE, 2025.

Hou-I Liu, Christine Wu, Jen-Hao Cheng, Wenhao Chai, Shian-Yun Wang, Gaowen Liu, Hugo Latapie, Jhih-Ciang Wu, Jenq-Neng Hwang, Hong-Han Shuai, et al. Monotakd: Teaching assistant knowledge distillation for monocular 3d object detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22266–22275. IEEE, 2025.

Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 7089–7097. IEEE, 2019.

Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3815–3824. IEEE, 2021.

Prajval Kumar Murali, Mohsen Kaboli, and Ravinder Dahiya. Intelligent in-vehicle interaction technologies. *Advanced Intelligent Systems*, 4(2):2100122, 2022.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971. IEEE, 2019.

Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Heterogeneous knowledge distillation using information flow modeling. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2336–2345. IEEE, 2020.

Hongyu Qu, Jianan Wei, Xiangbo Shu, Yazhou Yao, Wenguan Wang, and Jinhui Tang. Omnigaze: Reward-inspired generalizable gaze estimation in the wild. *Advances in Neural Information Processing Systems*, 38:105999–106026, 2026.

Mohammed Rakib and Arunkumar Bagavathi. G 2 d: Boosting multimodal learning with gradient-guided distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4059–4068. IEEE, 2025.

A Romero, N Ballas, S Ebrahimi Kahou, A Chassang, C Gatta, and Y Bengio. Fitnets: Hints for thin deep nets. arxiv 2014. *arXiv preprint arXiv:1412.6550*, 10.

Pierre Vuillecard and Jean-Marc Odobez. Enhancing 3d gaze estimation in the wild using weak supervision with gaze following labels. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13508–13518. IEEE, 2025.

Wenguan Wang, Jianbing Shen, Xingping Dong, Ali Borji, and Ruigang Yang. Inferring salient objects from human fixations. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):1913–1927, 2019.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19354–19363. IEEE, 2022.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 7130–7138. IEEE, 2017.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

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, 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 *European conference on computer vision*, pp. 365–381. Springer, 2020.

Long Zhao, Xi Peng, Yuxiao Chen, Mubbasir Kapadia, and Dimitris N Metaxas. Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6527–6536. IEEE, 2020.

A Appendix

A.1 Robustness to Full-frame Appearance Perturbations

Since the student operates directly on raw full-frame images, it may be affected by appearance variations unrelated to gaze. We therefore apply controlled appearance perturbations to the full-frame input while keeping the facial region unchanged. The perturbations include brightness variation, Gaussian noise, blur, and their combination.

Table 5 shows that all transfer strategies degrade under controlled appearance perturbations, with the combined perturbation causing the largest performance drop. Nevertheless, our method consistently maintains a lower angular error than the two transfer baselines. From the clean setting to the combined perturbation, the error of our method increases by 2.29°, compared with larger increases for Teacher Pseudo Supervision and feature-level alignment. These results suggest that the proposed transfer objectives encourage the full-frame student to rely more on gaze-relevant information and less on unstable appearance cues.

Table 5: Robustness to controlled full-frame appearance perturbations on LBW.

Perturbation	Pseudo	Feature Align.	Ours
None	17.66	16.37	**11.25**
Brightness	19.24	17.82	**12.18**
Gaussian Noise	20.13	18.46	**12.67**
Blur	19.51	17.94	**12.43**
Combined	22.06	20.31	**13.54**

A.2 Effect of Privileged Observation Diversity

Our framework extracts privileged gaze knowledge from multiple perturbed face observations rather than relying on a single teacher prediction. We therefore vary the number of privileged observations used for knowledge extraction while keeping the remaining training procedure unchanged. This experiment evaluates whether the diversity of privileged observations contributes to more stable knowledge transfer.

Table 6 shows that increasing the number of perturbed privileged observations generally improves the transfer performance. The improvement is more evident when increasing K from 1 to 4, while the gain becomes smaller with further increases. This result suggests that aggregating multiple privileged observations provides more stable gaze supervision than relying on a single teacher prediction, while a larger number of observations yields diminishing returns.

Table 6: Effect of the number of perturbed privileged observations on LBW.

Number of Views K	1	2	4	8	16
Angle ($^\circ$) $\downarrow$	13.18	12.47	11.79	11.41	**11.25**

A.3 Target Unlabeled Data Efficiency

We investigate how effectively the proposed framework exploits unlabeled target-domain data by varying the amount of training data while keeping the training protocol unchanged. No target gaze annotations are used during adaptation.

Table 7 shows that increasing the amount of unlabeled target-domain data consistently improves performance across all three driving conditions. On LBW, increasing the data from 20% to 100% reduces the angular error from 15.31° to 11.25°, corresponding to a relative reduction of 26.52%. The improvement is more pronounced when increasing the data from 20% to 60%, while the gains become smaller at larger data scales. Similar trends are observed on LBW-L and LBW-N. These results suggest that the proposed framework can effectively exploit additional unlabeled target observations, while the diminishing gains at larger data scales indicate that the benefit gradually saturates once sufficient target-domain diversity is observed.

Table 7: Effect of the amount of unlabeled target-domain data under different driving conditions. The last column reports the relative reduction in LBW angular error with respect to the 20% setting.

Data	LBW	LBW-L	LBW-N	Gain (%)
20%	15.31	18.42	18.87	–
40%	13.67	16.47	17.21	10.71
60%	12.76	15.32	15.94	16.66
80%	11.88	14.26	14.68	22.34
100%	**11.25**	**13.58**	**13.73**	**26.52**

A.4 LIMITATIONS

Despite the effectiveness of the proposed framework, several limitations remain. First, although face cropping is not required during inference, facial landmarks are still used during training to obtain the privileged face observations. This introduces an additional preprocessing dependency and assumes that sufficiently reliable face localization is available in the target training data. Second, the transferred knowledge is inherently influenced by the quality and distribution of the pretrained cropped-face gaze expert. Although the proposed perturbation-based confidence estimation reduces the impact of unstable target-domain predictions, substantial domain discrepancies may still limit the quality of privileged supervision. Third, our experiments are primarily conducted on the LBW driving dataset, and the challenging subsets are constructed according to image brightness. The results therefore provide limited evidence regarding generalization to substantially different camera configurations, driving environments, or other target domains. Extending the framework to broader cross-dataset and cross-sensor settings remains an important direction for future work.