LISA: Language-guided Interference-aware Spatial-Frequency Attention for Driver Gaze Estimation

Jun Ma[1,2], **Zhenye Yang**[1,2], **Ruichen Zhou**[1,2], **Pei Zhang**[3], **Huan Li**[4] and **Jinpeng Chen**[1,2*]
[1]School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications
[2]Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education
[3]School of Electrical Engineering, Guangxi University
[4]Zhejiang University
{majun, yangzhenye, 2025010344}@bupt.edu.cn, pzhang@gxu.edu.cn, lihuan.cs@zju.edu.cn, jpchen@bupt.edu.cn

Abstract

Driver gaze estimation serves as a fundamental metric for evaluating driver attentiveness in modern monitoring systems. Beyond being vulnerable to sudden lighting changes and sensor noise, spatial-domain models struggle to disentangle authentic gaze cues from irrelevant visual attributes. In this paper, we propose LISA, a **L**anguage-guided **I**nterference-aware **S**patial-Frequency **A**ttention framework that combines frequency-domain priors with vision-language knowledge. Observing that the amplitude spectrum remains relatively stable even under spatial perturbations, we design a dual-domain fusion mechanism. It integrates stable low-frequency semantics into high-frequency details, employing spatial attention to precisely target ocular regions. To reduce semantic ambiguity, we also introduce a training-time disentanglement strategy. Using a frozen CLIP encoder and orthogonal regularization, we explicitly separate gaze features from appearance interference. Experiments on two benchmarks show that LISA achieves state-of-the-art performance, with significantly improved robustness against occlusions and lighting variations. The code repository is available at https://github.com/Mason-bupt/LISA.

1 Introduction

Driver distraction and fatigue are among the major contributors to global traffic accidents, motivating the deployment of effective driver monitoring systems (DMS) to improve road safety [Rangesh *et al.*, 2020; Guo *et al.*, 2024]. Among various behavioral cues, gaze direction is a key indicator of driver cognitive state and attention concentration [Hu *et al.*, 2024]. Although early solutions relied on supportive devices such as eye tracking glasses or specialized infrared sensors, the field has clearly shifted towards using standard RGB cameras for appearance-based gaze estimation [Liu *et al.*, 2021; Cheng *et al.*, 2022]. These data-driven approaches, powered by Convolutional Neural Networks (CNNs) and Transformers, offer a non-invasive and cost-effective solution suitable for deployment in modern smart cockpits.

However, the transition of these models from a controlled environment to a real driving scenario poses many challenges. The visual environment in the vehicle is inherently unstable and drivers frequently encounter extreme lighting conditions, from dazzling sunlight to dimly lit tunnels [Li *et al.*, 2024]. This complexity is further compounded by various physical attributes and occluding accessories, such as sunglasses and masks [Muhammad *et al.*, 2020].

In the presence of such visual instability, mainstream methods primarily employ deep convolutional backbone networks or visual transformers to extract high-level semantic features. These methods rely on the aggregate of local pixel patterns or patch-based markers to regress gaze vectors [Kellnhofer *et al.*, 2019; Cheng and Lu, 2022; Cheng *et al.*, 2024b]. However, a critical limitation persists: these architectures operate primarily in the spatial domain. Consequently, they are inherently sensitive to pixel-level interference. When visual conditions change, the spatial distribution of key ocular features shifts unpredictably, causing the models to produce jittery or erroneous predictions [Zhang *et al.*, 2015; Cheng *et al.*, 2024a]. Moreover, without explicit disentanglement mechanisms, they tend to overfit irrelevant appearance attributes (e.g., associating "sunglasses" with a fixed gaze bias) rather than learning the true gaze feature [Cheng *et al.*, 2022; Kim *et al.*, 2024].

This vulnerability to environmental shifts exposes the basic limitation of the current CNN-based paradigm, which we define as **Spatial Misalignment**. As illustrated in Figure 1, spatial features are highly sensitive to disturbances at the pixel-level. Under the condition of weak light or noise (Condition B), the feature distribution of the human eye region is significantly more degraded than that under the clear condition (Condition A), which makes it difficult for the model to capture subtle fixation clues.

Although the spatial pixel values fluctuate greatly, we observe that the overall structure of the face remains basically unchanged in the frequency domain. As shown in the fre-

*Corresponding author.

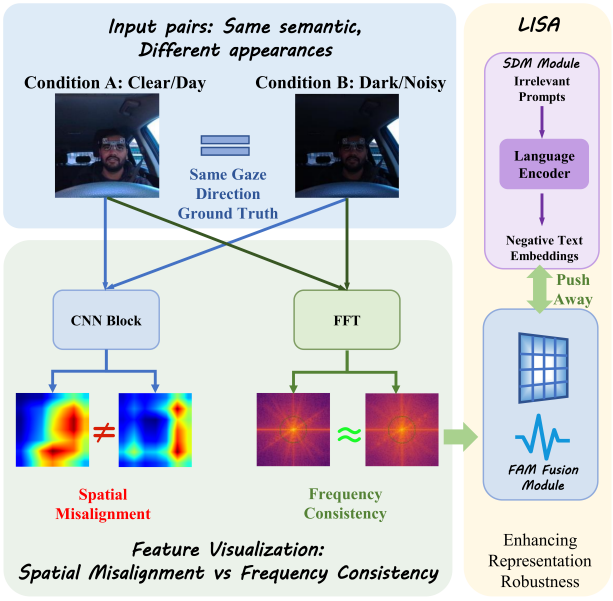

Figure 1: Visualizing the robustness gap. **Left:** Spatial features fail to align under lighting changes, while frequency spectra remain stable. **Right:** LISA enhances robustness by injecting frequency priors and employing a language-guided "Push Away" mechanism to suppress irrelevant semantic attributes.

quency path of Figure 1, the amplitude spectrum shows **Frequency Consistency**, which can maintain a stable structural characteristic [Chen *et al.*, 2021] even when the spatial image is corrupted. This observation prompted us to explore the frequency-domain prior as a stabilizer for gaze estimation.

To address the dual challenges of environmental instability and appearance entanglement, we propose LISA (**L**anguage-guided **I**nterference-aware **S**patial-Frequency **A**ttention), a robust framework for unconstrained driver gaze estimation. Our method is based on two core components. First, to mitigate environmental shifts, we introduce the FAM Fusion (Frequency-Attention Modulated Fusion) module within a ResNet-18 backbone [He *et al.*, 2016]. Unlike traditional spatial fusion, FAM Fusion incorporates a Spectral Injection Block that converts features into the frequency domain through the Fast Fourier Transform (FFT). This mechanism can isolate stable low-frequency semantic components and adaptively inject them into high-frequency details, thereby maintaining frequency consistency when spatial features are affected by lighting or noise. As a supplement to this spectrum alignment, we integrate a Spatial Saliency Gating mechanism. This spatial attention module explicitly reweights the feature map to highlight key eye regions, ensuring the precise localization of visual cues. Second, to resolve appearance entanglement, we propose a Semantic Disentanglement Module (SDM). We implemented a feature separation training strategy using cross-modal knowledge from visual language models (specifically a frozen CLIP encoder [Radford *et al.*, 2021]). As depicted by the **Push Away** mechanism in Figure 1, this forces the model to minimize the similarity between gaze features and text embeddings of common distractors (e.g., "a driver wearing sunglasses"), resulting in pure gaze representations that are orthogonal to appearance noise.

Our contributions are as follows:

- We propose LISA, a Language-guided Interference-aware Spatial-Frequency Attention framework, which is anchored in the concept of frequency consistency to tackle persistent spatial misalignment bottleneck caused by environmental shifts.
- Our method integrates the FAM Fusion module to obtain stable spectral semantics and the SDM module to filter appearance-based ambiguity, achieving excellent performance in multiple benchmark tests.

2 Related Work

2.1 Gaze Estimation

Monitoring driver gaze direction can effectively reduce the incidence of road accidents [Li and Li, 2015; Zhang *et al.*, 2017; Wan *et al.*, 2021b; Wan *et al.*, 2021a]. Since gaze direction is difficult to measure directly, previous work has transformed gaze estimation into a sophisticated 3D representation [Xiao *et al.*, 2025; Kawana *et al.*, 2025]. Furthermore, certain methodologies incorporate additional features, such as synthetic 3D annotations [Bao *et al.*, 2025], 3D priors with weak supervision [Cheng *et al.*, 2025; Vuillecard and Odobez, 2025], and large-scale encoders for target estimation [Ryan *et al.*, 2025]. Generally, the incorporation of additional information can significantly improve the model's predictive accuracy.

However, gaze is inherently a continuous spatial behavior, and capturing the temporal dynamics can enhance the accuracy of the estimation. Certain methodologies employ LSTMs [Kellnhofer *et al.*, 2019] and attention mechanisms [Jindal *et al.*, 2024] to implicitly model temporal relevance. [Yu *et al.*, 2025] and [Mazzamuto *et al.*, 2025] proposed approaches that capture gaze dynamics, thereby enhancing video gaze estimation capabilities. The construction of a gaze estimation model encounters challenges such as disentangling authentic gaze cues from visual clutter. Additionally, [Du and Lan, 2022] and [Yin *et al.*, 2024] proposed frameworks based on frequency domains and prompts.

2.2 Gaze Estimation for Driver

While invasive active sensors have been extensively researched, there has been a recent shift towards non-intrusive vision-based systems. Existing studies often approximated gaze by estimating sparse zones [Yang *et al.*, 2021; Hu *et al.*, 2024]. Consequently, with the advent of modern benchmarks, researchers begin to explore how to predict accurate 3D direction continuously. [Kasahara *et al.*, 2022] and [Cheng *et al.*, 2024b] proposed frameworks based on self-supervision and dual-stream Transformers to capture complex line-of-sight patterns. In particular, FIFA [Hu *et al.*, 2025] introduces a fine-grained inter-frame attention module designed to track subtle pupil movements over time. However, the susceptibility of spatial domain methods to environmental changes restricts their applicability.

3 Methodology

In this section, we provide a comprehensive delineation of the proposed LISA framework. In detail, first, we introduce the problem formulation. Second, we provide an overview of the architecture. Next, we detail the core FAM Fusion module and the SDM. Finally, we describe the Regression Head and optimization objectives.

3.1 Problem Formulation

Specifically, we utilize a single RGB input image $\mathbf{X} \in \mathbb{R}^{3\times H\times W}$ depicting a driver to estimate the gaze direction. Considering that ground truth annotations in standard gaze datasets are typically provided as 3D unit vectors $\mathbf{v} = (x, y, z) \in \mathbb{R}^3$, we formulate the regression task using 2D spherical coordinates, specifically predicting the angles of yaw (ϕ) and pitch (θ), i.e., $\mathbf{g} = (\phi, \theta) \in \mathbb{R}^2$. Since the 2D pitch and yaw angles are sufficient to uniquely describe the gaze direction within the head coordinate system, we adopt this 2D representation. Additionally, in order to be generally more stable and easier for the model to learn compared to a 3D vector constrained to the unit sphere, we regress a 2D continuous variable. This conversion between the 3D unit vector $\mathbf{v}$ and the 2D angles $\mathbf{g}$ follows standard spherical coordinate transformations. As a result, it allows for seamless integration with 3D evaluation metrics.

3.2 Framework Overview

As illustrated in Figure 2, the proposed LISA framework is built on a lightweight ResNet backbone. Specifically, the backbone extracts multi-scale feature maps: Detail Features $\mathbf{F}_{detail}$ from the shallow layer, which preserves high-frequency spatial cues. Meanwhile, the deep features $\mathbf{F}_{guide}$ of the deep layer contain a rich semantic context. The aligned features $\mathbf{F}'_{detail}$ and $\mathbf{F}'_{guide}$ are fed into the FAM Fusion module, where they are fused via spectral and spatial mechanisms. The fused feature is then processed by a Regression Head to predict the gaze vector $\mathbf{g}$. To facilitate the learned features to be robust to appearance interference, we propose an auxiliary Semantic Disentanglement Module, which leverages a frozen CLIP encoder to regularize the latent space.

3.3 Feature Alignment

In order to align different scales and channel dimensions of the multi-level features extracted by the backbone prior to the core fusion process, we employ a Feature Alignment block to unify these properties. Consequently, the detail and guide features can be processed effectively in subsequent modules. We denote the aligned features by $\mathbf{F}'_{detail}$ and $\mathbf{F}'_{guide}$.

The alignment process is formulated as follows.

$$\mathbf{F}'_{detail} = \phi_{align}(\mathbf{F}_{detail}) \tag{1}$$

$$\mathbf{F}'_{guide} = \text{Upsample}(\phi_{align}(\mathbf{F}_{guide})) \tag{2}$$

Specifically, ϕ_{align} includes a 1×1 convolution followed by Batch Normalization and GELU activation; the projection maps both feature streams to a unified channel dimension C. Additionally, in order to match the spatial resolution of the detail features $\mathbf{F}_{detail}$, we apply bilinear upsampling to the deep features $\mathbf{F}_{guide}$.

3.4 Frequency-Attention Modulated Fusion (FAM Fusion)

Traditional methods still suffer from "Spatial Misalignment", owing to their heavy reliance on unstable spatial pixels. As the environment changes, the issue worsens markedly. To bridge this gap, we propose a Frequency-Attention Modulated Fusion, called FAM Fusion. To facilitate the network in guaranteeing frequency consistency, we design a Spectral Injection Block. Furthermore, a Spatial Saliency Gating mechanism is integrated to sharpen local attention, thereby specifically improving focus.

Spectral Injection Block

We introduce a Spectral Injection Block aimed at extracting semantic priors from the deep branch and replacing the low-frequency components of the detail branch.

FFT Transformation. We map spatial features to the frequency domain using a 2D real-to-complex Fast Fourier Transform (rFFT), denoted as $\mathcal{F}_r$. Specifically, $\mathcal{F}_r$ encodes the aligned detail and guide features $\mathbf{F}'_{detail}$ and $\mathbf{F}'_{guide}$ into complex-valued spectral matrices $\mathbf{Y}_{detail}$ and $\mathbf{Y}_{guide}$:

$$\mathbf{Y}_{detail} = \mathcal{F}_r(\mathbf{F}'_{detail}) \tag{3}$$

$$\mathbf{Y}_{guide} = \mathcal{F}_r(\mathbf{F}'_{guide}) \tag{4}$$

where $\mathcal{F}_r$ denotes the rFFT operation, and $\mathbf{Y} \in \mathbb{C}^{C\times H\times(\lfloor W/2\rfloor+1)}$ represents the complex-valued spectral matrix.

Adaptive Spectral Mixing. We define a binary mask matrix $\mathbf{M}$ to isolate the low-frequency region determined by a ratio γ. A learnable scalar parameter $\alpha \in (0, 1)$ is used as an adaptive weight for spectral mixing. The fused spectrum $\mathbf{Y}_{fused}$ is obtained as follows:

$$\mathbf{Y}_{fused} = \mathbf{Y}_{detail} \cdot (1 - \alpha \cdot \mathbf{M}) + \mathbf{Y}_{guide} \cdot (\alpha \cdot \mathbf{M}) \tag{5}$$

where $\mathbf{M}$ is a binary mask matrix isolating the low-frequency region determined by a ratio γ, and $\alpha \in (0, 1)$ is a scalar parameter that can be learned by constraining a sigmoid function. This framework extracts semantic priors from the deep stream and assigns weights to them, thereby preserving high-frequency edge information from $\mathbf{Y}_{detail}$ and substituting its low-frequency components with $\mathbf{Y}_{guide}$.

Finally, the inverse rFFT operation recovers the spatial feature $\mathbf{F}_{freq}$:

$$\mathbf{F}_{freq} = \mathcal{F}_r^{-1}(\mathbf{Y}_{fused}) \tag{6}$$

where $\mathcal{F}_r^{-1}$ denotes the inverse rFFT operation.

Spatial Saliency Gating

The representation of spatial features is influenced by the coupling factors of global structural stability and background noise. Therefore, we employ Spatial Saliency Gating to capture the ocular region information, thereby suppressing background noise.

Specifically, a lightweight network is employed to extract attention features from the deep features $\mathbf{F}'_{guide}$. We generate a spatial attention matrix $\mathbf{A} \in \mathbb{R}^{1\times H\times W}$ as follows:

$$\mathbf{A} = \sigma(\text{Conv}_2(\delta(\text{Conv}_1(\mathbf{F}'_{guide})))) \tag{7}$$

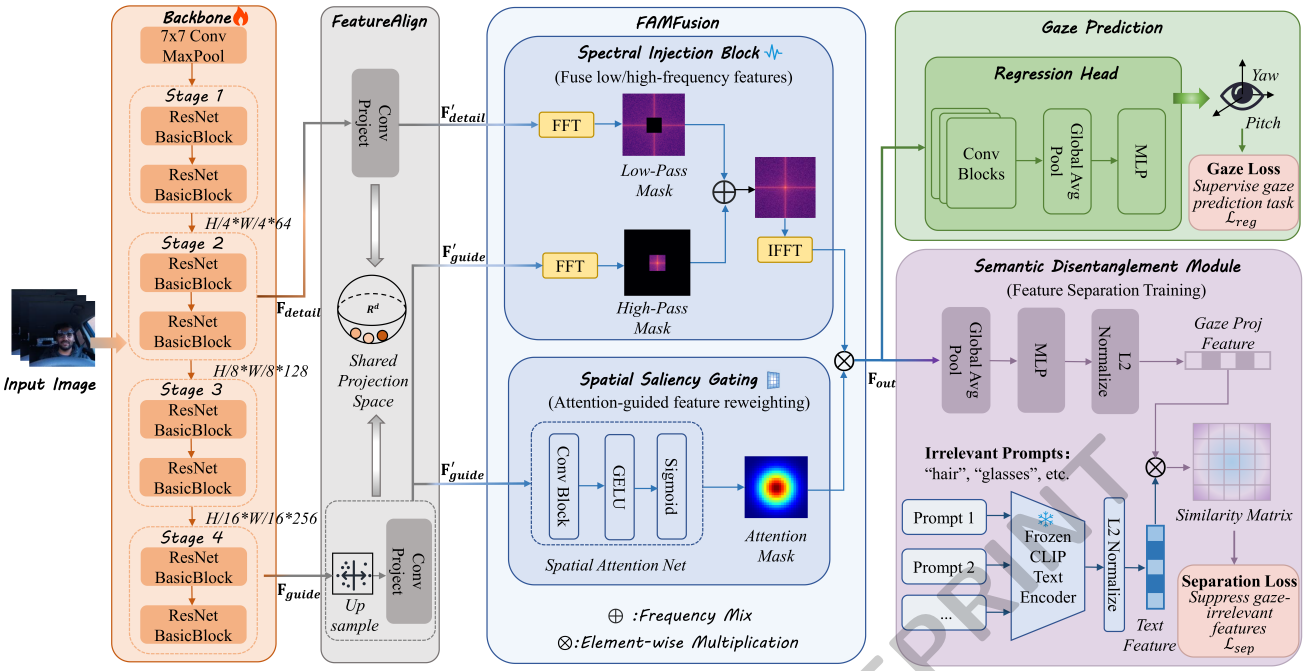

Figure 2: Overview of the LISA framework. The architecture consists of a ResNet-18 backbone extracting multi-scale features, a Feature Alignment module for dimension unification, the core FAM Fusion module, and the Semantic Disentanglement Module.

where $\mathbf{A} \in \mathbb{R}^{1\times H\times W}$ is the attention map, σ is the Sigmoid function, δ is the GELU activation, and $\text{Conv}_{1,2}$ are convolutional layers.

Finally, the output features are computed as follows:

$$\mathbf{F}_{out} = \mathbf{F}_{freq} \cdot (\mathbf{A} + \epsilon) \tag{8}$$

Here, ϵ indicates a residual scalar bias term used to stabilize training in its early stages.

3.5 Semantic Disentanglement Module (SDM)

In order to resolve the Semantic Ambiguity where gaze features are entangled with appearance attributes, we need to ensure that the learned representations contain orthogonal regularization. To achieve this goal, we utilize the rich semantic knowledge of a pre-trained Vision-Language Model and introduce a Semantic Disentanglement Module (SDM).

Text-Driven Negative Anchors. Specifically, for a given set of "negative" semantic anchors that represent gaze-irrelevant appearance factors, where T is the prompt pool, the diverse appearance attributes are expressed as a set of N prompts $T = \{t_1, t_2, \ldots, t_N\}$ through natural language templates. In detail, we freeze the CLIP Text Encoder and encode these prompts to obtain a semantic embedding matrix $\mathbf{E}_{text} \in \mathbb{R}^{N\times D}$.

Feature Projection. To elaborate, we project the fused spatial feature $\mathbf{f}_{gaze}$ to an intermediate representation $\mathbf{e}_{gaze}$ with a projection head ϕ_{proj}. A Linear layer is then employed to project these feature vectors into a shared CLIP latent space:

$$\mathbf{e}_{gaze} = \phi_{proj}(\mathbf{f}_{gaze}) \in \mathbb{R}^D \tag{9}$$

where $\mathbf{e}_{gaze}$ is the projected visual embedding.

Orthogonal Regularization. Since the main objective is to ensure that the gaze representation $\mathbf{e}_{gaze}$ contains no information related to the appearance attributes defined in $\mathbf{E}_{text}$, we hope that the optimizer can minimize any correlation between the gaze feature and the appearance text embeddings. By calculating the cosine similarity between the projected gaze feature and each text embedding and applying the mean operation, we obtain the separation loss $\mathcal{L}_{sep}$, which encodes the correlation between the gaze feature and the appearance text embeddings. The separation loss is defined as:

$$\mathcal{L}_{sep} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\mathbf{e}_{gaze}\cdot\mathbf{E}_{text,i}}{\|\mathbf{e}_{gaze}\|\|\mathbf{E}_{text,i}\|}\right| \tag{10}$$

where $|\cdot|$ denotes the absolute value operation, ensuring that the optimizer minimizes any correlation (positive or negative) between the gaze feature and the appearance text embeddings $\mathbf{E}_{text,i}$.

3.6 Regression Head and Loss Function

Coordinate-aware Regression. Specifically, for a given pair of generated normalized coordinate maps $\mathbf{C}_x, \mathbf{C}_y \in \mathbb{R}^{H\times W}$, where (h, w) is the spatial location, the position information is represented as a value defined as:

$$\mathbf{C}_x(h,w) = \frac{2w}{W-1} - 1 \tag{11}$$

$$\mathbf{C}_y(h,w) = \frac{2h}{H-1} - 1 \tag{12}$$

Normalized spatial coordinates are concatenated according to channel dimension, resulting in a complete feature map

$\mathbf{F}_{coord}$:

$$\mathbf{F}_{coord} = \text{Concat}(\mathbf{F}_{out}, \mathbf{C}_x, \mathbf{C}_y) \tag{13}$$

The regression head consists of a Multi-Layer Perceptron (MLP) aimed at predicting final gaze angles from the augmented feature $\mathbf{F}_{coord}$ and pooling it across spatial locations:

$$\hat{\mathbf{g}} = \Phi_{reg}(\text{Pool}(\mathbf{F}_{coord})) \tag{14}$$

where Φ_{reg} denotes the MLP regression network, and $\hat{\mathbf{g}} = (\hat{\phi}, \hat{\theta})$ represents the predicted yaw and pitch angles.

Total Objective. We use the following composite loss function to minimize our total objective:

$$\mathcal{L}_{total} = \mathcal{L}_{reg} + \lambda_{sep}\mathcal{L}_{sep} \tag{15}$$

To strictly constrain the geometric consistency of the gaze direction, a regression loss $\mathcal{L}_{reg}$ is used to combine the Smooth L1 loss and the Angular Loss, thus constraining the geometric consistency of the predicted 3D vector:

$$\mathcal{L}_{reg} = \mathcal{L}_{L1}(\mathbf{g}, \hat{\mathbf{g}}) + \lambda_{ang}\mathcal{L}_{ang}(\mathbf{v}, \hat{\mathbf{v}}) \tag{16}$$

where $\mathcal{L}_{L1}$ denotes the loss in pitch and yaw angles, and $\mathcal{L}_{ang}$ represents the angular error in degrees between the predicted 3D vector $\hat{\mathbf{v}}$ and the ground truth vector $\mathbf{v}$.

4 Experiments

4.1 Datasets

We used the following datasets in our experiments: LBW [Kasahara *et al.*, 2022] and IVGaze [Cheng *et al.*, 2024b].

LBW is a large-scale dataset that includes driver facial images, road-facing scene images, and 3D gaze directions from head-mounted eye trackers. Facial landmarks are used to crop facial images with a fixed resolution size. The dataset contains 28 subjects. We divided the dataset into two subsets based on subjects, with subjects with IDs 1–22 as a training set and the rest as a test set.

The IV dataset comprises 44,703 images in 125 subjects. It encompasses a rich set of ground truth data, including head pose and gaze direction. We adhere to the dataset division strategy outlined in [Cheng *et al.*, 2024b] and implement a three-fold cross-validation approach.

4.2 Implementation Details

The proposed method is implemented using PyTorch and trained for 100 epochs on an NVIDIA RTX 4090 GPU. A batch size of 64 is used for training, and the learning rate is set to 1×10^{-4}. Optimization is performed using the AdamW optimizer with a weight decay of 5×10^{-4}.

4.3 Comparison with SOTA Methods

We compare our approach with the following baselines: (i) ResNet-18 [He *et al.*, 2016], (ii) Gaze360 [Kellnhofer *et al.*, 2019], (iii) XGaze [Zhang *et al.*, 2020], (iv) GazeTR [Cheng and Lu, 2022], (v) GazePTR [Cheng *et al.*, 2024b], (vi) STAGE [Jindal *et al.*, 2024] and (vii) FIFA [Hu *et al.*, 2025]. All the compared baselines rely on CNN- or Transformer-based (or hybrid) feature extractors to learn appearance representations for gaze estimation.

Method	Error		Specification	
	IV	LBW	FLOPs	#Param.
Gaze360	8.23°	7.12°	12.78G	14.60M
ResNet-18	8.02°	6.41°	3.65G	11.25M
GazeTR	7.33°	6.17°	3.68G	11.39M
XGaze	7.35°	6.23°	8.26G	23.64M
GazePTR	7.09°	6.05°	3.69G	11.52M
STAGE	7.34°	6.12°	3.67G	11.32M
FIFA	7.29°	5.84°	**2.70G**	5.87M
Ours	$\mathbf{7.05}^\circ$	$\mathbf{5.39}^\circ$	5.65G	**4.39M**

Table 1: Performance comparison. Best results are marked in **bold**.

The mean angle error is used to evaluate the accuracy of the model prediction, with lower values representing better performing methods.

In Table 1, we present a comparative analysis of the performance of our method against state-of-the-art techniques. Initially, in the IV dataset, our method achieves an improvement with an error of 7.05° representing a reduction of 0.97° relative to the ResNet-18's error of 8.02°. The performance of our proposed method also surpasses the recent hybrid architecture FIFA (7.29°) by a margin of 0.24°. For the LBW dataset, our method similarly exhibits superior performance, achieving an error of 5.39°, which signifies a decrease of 1.02° from ResNet-18's error of 6.41°. Compared to FIFA (5.84°), our proposed method enhances performance by 7.7%, demonstrating that our method outperforms current methods in terms of accuracy to ensure high precision and robustness in complex driving scenarios challenged by various visual interferences.

For FLOPs, our method requires 5.65G operations to support frequency-domain modeling, representing a moderate computational cost for improved accuracy. Moreover, our method encompasses only 4.39M parameters, a significantly smaller number compared to Gaze360's 14.6M and XGaze's 23.64M, rendering the model more suitable for deployment in resource-limited settings.

4.4 Visualization of Prediction Results

When driving a vehicle, the gaze direction is usually fixed in a few areas. Figure 3 highlights the utilization of the average precision within the pitch and yaw dimensions to assess the model's performance at varying levels of permitted error. The MAE heatmap (Left) shows an expansive high-precision region, maintaining consistently low error rates where the pitch range is $[-5^\circ, 15^\circ]$ and the yaw range is $[-40^\circ, 40^\circ]$. This outcome ensures that a larger number of samples maintains prediction accuracy within the designated error limits, which is particularly beneficial for driver monitoring tasks. By cross-referencing with the sample density map (Right), we observe a matching accuracy in areas where samples are concentrated, and the proposed method can obtain a more accurate gaze estimation. Moreover, the model's performance exhibits a smooth error gradient rather than abrupt performance drops, suggesting a higher concentration of samples within these bounds. In general, these results demonstrate

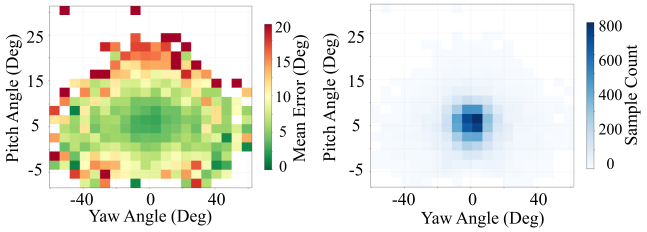

Figure 3: Error and sample distribution in gaze angle space. The x-axis denotes yaw angle, and the y-axis denotes pitch angle.

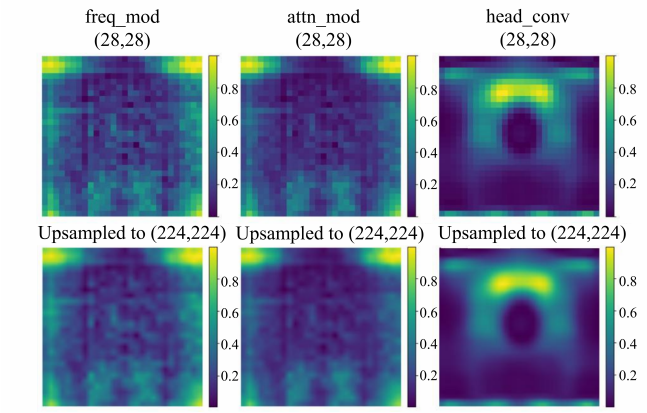

Figure 4: Visualization of multi-resolution feature evolution.

that LISA-based representations are effective and can be easily transferred to most real-world applications.

4.5 Embedding Visualization

To substantiate the efficacy of the proposed internal components, we present a visualization of the feature maps of the test dataset. Figure 4 illustrates the progressive evolution of feature maps in the LBW dataset through the *Spectral Injection Block*, *Spatial Saliency Gating* and the final Regression Head output. For the feature distributions of the initial activations in the evaluation datasets, it can be seen that the features exhibit distributed, texture-rich patterns, confirming that our frequency-aware fusion explicitly preserves fine-grained structural details. Crucially, we observe that the final output embeddings exhibit more refined distributions rather than a simple localized hotspot. This transformation demonstrates that the head successfully integrates spatial coordinates with global orientation cues to bridge the gap between abstract features and continuous gaze regression.

To rigorously validate the generalization capabilities of the model, we employ t-SNE to project high-dimensional feature embeddings into a 2D manifold, as visualized in Figure 5.

We first examine the *Prediction Error Distribution* (Left), which reveals a highly structured topology regarding the placement of correct and incorrect samples. Rather than a random scatter, the manifold exhibits a clear "core-periphery" organization: samples with high estimation precision (dense green regions) constitute the primary body of the distribution, while high-error samples (red/yellow) are largely marginalized to the sparse peripheries. This indicates that the model has successfully learned a robust mapping for the vast majority of the gaze distribution, effectively pushing failure cases—likely corresponding to extreme angles—to the geometric boundaries of the feature space.

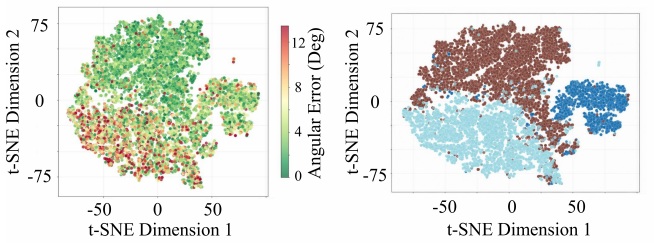

Figure 5: t-SNE visualization; axes denote t-SNE dimensions, with colors indicating error and subject ID, respectively.

Turning to the *Subject ID Clustering* (Right), we observe that the embeddings naturally segregate into distinct clusters based on individual identities, confirming that the model preserves inherent appearance variations such as facial morphology. Crucially, by cross-referencing these two views, we find that the favorable error topology observed in the Left plot is consistently replicated within *each* subject-specific cluster. Specifically, every identity cluster exhibits the same stable structure: a dense, accurate core surrounded by sparse, higher-error outliers. This structural consistency demonstrates that while the model encodes subject-specific differences (hence the separation), its gaze estimation mechanism remains fundamentally stable and unbiased across different individuals.

4.6 Robustness Analysis Against Occlusions and Environmental Factors

We perform a 3-fold cross-validation on 44,703 IV dataset samples to evaluate LISA under unconstrained conditions. Table 2 compares LISA with GazePTR on accessories, occlusions, and environmental shifts. LISA achieves a lower mean error of $\mathbf{7.05}^\circ$ and a reduced standard deviation ($\mathbf{5.40}^\circ$ vs. 5.61°). These results demonstrate superior stability and robustness across diverse driving scenarios.

Resilience to Accessories and Visual Clutter. A broad analysis of the "w/ Accessories" category reveals the model's exceptional stability amidst visual distractions. Despite the introduction of diverse facial ornaments in 30,720 samples, our method maintains a competitive error of 7.29°. Performance degradation relative to the clean baseline ("w/o Accessories") is limited to 0.76°. This minimal impact confirms that our feature extractor effectively disentangles gaze-relevant cues from extraneous appearance variations, preventing the model from overfitting to irrelevant facial ornaments.

Robustness Against Specific Occlusions. The model demonstrates remarkable resilience to common interferences. In scenarios involving eyewear ("w/ Glasses"), which often introduce specular reflections, our method achieves an error of 7.20°, outperforming GazePTR (7.22°) with a marginal impact of only 0.44° compared to subjects without glasses.

Category	Count	Mean Error (↓)		Std. Dev.		Accuracy ($< 8°$ ↑)	
		GazePTR	**Ours**	GazePTR	**Ours**	GazePTR	**Ours**
Overall	44,703	$7.09°$	$\mathbf{7.05}°$	$5.61°$	$\mathbf{5.40}°$	66.3%	$\mathbf{66.8}\%$
Impact of Accessories							
w/ Accessories	30,720	$7.30°$	$\mathbf{7.29}°$	$5.81°$	$\mathbf{5.49}°$	64.8%	$\mathbf{65.6}\%$
w/o Accessories	13,983	$6.66°$	$\mathbf{6.53}°$	$5.19°$	$\mathbf{5.11}°$	69.1%	$\mathbf{69.7}\%$
Impact of Occlusions							
w/ Glasses	27,774	$7.22°$	$\mathbf{7.20}°$	$5.62°$	$\mathbf{5.38}°$	65.3%	$\mathbf{65.8}\%$
w/o Glasses	16,929	$6.88°$	$\mathbf{6.76}°$	$5.47°$	$\mathbf{5.34}°$	67.3%	$\mathbf{68.5}\%$
w/ Mask	8,331	$7.81°$	$\mathbf{7.54}°$	$6.31°$	$\mathbf{5.93}°$	63.2%	$\mathbf{65.2}\%$
w/o Mask	36,372	$6.92°$	$\mathbf{6.92}°$	$5.41°$	$\mathbf{5.25}°$	66.9%	$\mathbf{67.2}\%$
Impact of Environment							
Outdoor	40,865	$7.14°$	$\mathbf{7.12}°$	$5.66°$	$\mathbf{5.44}°$	65.8%	$\mathbf{66.3}\%$
Indoor	3,838	$\mathbf{6.00}°$	$6.06°$	$4.75°$	$\mathbf{4.48}°$	$\mathbf{73.4}\%$	73.1%

Table 2: Comparative robustness analysis between GazePTR model and our proposed LISA framework. Best results are marked in **bold**.

Method	IV Mean	LBW Mean
ResNet-18	$8.02°$	$6.41°$
w/o Spectral Injection Block	$7.41°$	$6.33°$
w/o Spatial Saliency Gating	$7.25°$	$5.88°$
w/o SDM	$7.24°$	$5.54°$
Ours	$\mathbf{7.05}°$	$\mathbf{5.39}°$

Table 3: Ablation study results (Mean Error).

This validates the efficacy of our Frequency Modulation module, which suppresses high-frequency glare artifacts. Notably, the advantage of our framework becomes more pronounced in severe occlusion scenarios such as facial masks ("w/ Mask"). Although GazePTR suffers a significant performance drop to $7.81°$, our method maintains high precision at $\mathbf{7.54}°$—a substantial improvement of $\mathbf{0.27}°$. Furthermore, our method achieves a much lower standard deviation ($5.93°$ vs. $6.31°$), proving that the Attention Modulation mechanism successfully redirects focus to exposed ocular regions to compensate for the loss of lower-facial context.

Environmental Adaptability. The environmental analysis highlights the model's robustness to illumination changes. In the "Outdoor" setting, which accounts for the majority of wild samples, our model achieves an error of $7.12°$, consistently outperforming GazePTR ($7.14°$). While GazePTR shows a slight advantage in mean error under ideal "Indoor" conditions ($6.00°$ vs. $6.06°$), our method exhibits superior stability with a significantly lower standard deviation ($4.48°$ vs. $4.75°$). This indicates that the LISA framework delivers more consistent predictions in varying lighting conditions, ensuring reliability whether it is deployed in controlled indoor environments or harsh outdoor settings.

4.7 Ablation Study

To systematically investigate the contribution of each component within our LISA framework, we conducted a comprehensive ablation study, as summarized in Table 3. While all ablated variants consistently outperform the ResNet-18 baseline on IV/LBW ($8.02°$/$6.41°$), the full model achieves optimal performance ($\mathbf{7.05}°$/$\mathbf{5.39}°$), indicating that these modules work synergistically rather than independently.

Specifically, removing the *Spectral Injection Block* results in the most significant degradation, particularly in cross-subject generalization (LBW error rises sharply to $6.33°$). This confirms that frequency-domain interactions are indispensable for learning domain-invariant representations, effectively filtering out subject-specific high-frequency noise while retaining stable structural semantics. Similarly, excluding *Spatial Saliency Gating* leads to a notable increase in error (up to $7.25°$ on IV), validating its critical role in suppressing background clutter and focusing the model on informative ocular regions. Finally, the exclusion of the *Semantic Disentanglement Module (SDM)* also incurs a performance penalty. This result validates the necessity of our CLIP-guided orthogonal regularization, thereby preventing the model from learning spurious correlations.

5 Conclusion

In this paper, we introduce LISA, a framework designed to address persistent bottlenecks of spatial misalignment and semantic ambiguity in driver gaze estimation. By synergizing frequency-domain consistency via the FAM Fusion module with CLIP-guided semantic disentanglement, our architecture effectively stabilizes features against environmental perturbations and explicitly decouples authentic gaze cues from appearance interference, such as accessories and glare. Extensive experiments on the IV and LBW benchmarks validate that LISA achieves state-of-the-art accuracy while maintaining computational efficiency, demonstrating superior resilience to severe occlusions and lighting changes compared to existing baselines. Future work will explore the extension of this paradigm to the temporal domain and event-based camera modalities to further improve robustness under extreme dynamic conditions.

Acknowledgements

This work was supported in part by the Beijing Natural Science Foundation(Grant No.L233034, No.L253004), in part by the National Natural Science Foundation of China (No.62572075) and in part by Fundamental Research Funds for the Beijing University of Posts and Telecommunications (No.2025TSQY01).

References

[Bao *et al.*, 2025] Yiwei Bao, Zhiming Wang, and Feng Lu. GazeGene: Large-scale synthetic gaze dataset with 3d eyeball annotations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18749–18759, 2025.

[Chen *et al.*, 2021] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 458–467, 2021.

[Cheng and Lu, 2022] Yihua Cheng and Feng Lu. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3341–3347, 2022.

[Cheng *et al.*, 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, pages 436–443, 2022.

[Cheng *et al.*, 2024a] 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.

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

[Cheng *et al.*, 2025] Yihua Cheng, Hengfei Wang, Zhongqun Zhang, Yang Yue, Boeun Kim, Feng Lu, and Hyung Jin Chang. 3D prior is all you need: Cross-task few-shot 2d gaze estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23891–23900, 2025.

[Du and Lan, 2022] Lingyu Du and Guohao Lan. Freegaze: resource-efficient gaze estimation via frequency domain contrastive learning. *arXiv preprint arXiv:2209.06692*, 2022.

[Guo *et al.*, 2024] Zizheng Guo, Qing Liu, Lin Zhang, Zhenning Li, and Guofa Li. L-tla: A lightweight driver distraction detection method based on three-level attention mechanisms. *IEEE Transactions on Reliability*, 73(4):1731–1742, 2024.

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

[Hu *et al.*, 2024] Zhongxu Hu, Yuxin Cai, Qinghua Li, Kui Su, and Chen Lv. Context-aware driver attention estimation using multi-hierarchy saliency fusion with gaze tracking. *IEEE Transactions on Intelligent Transportation Systems*, 25(8):8602–8614, 2024.

[Hu *et al.*, 2025] 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*, pages 18760–18769, 2025.

[Jindal *et al.*, 2024] Swati Jindal, Mohit Yadav, and Roberto Manduchi. Spatio-temporal attention and gaussian processes for personalized video gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 604–614, 2024.

[Kasahara *et al.*, 2022] 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*, pages 126–142, 2022.

[Kawana *et al.*, 2025] Yuki Kawana, Shintaro Shiba, Quan Kong, and Norimasa Kobori. GA3CE: Unconstrained 3d gaze estimation with gaze-aware 3d context encoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3081–3090, 2025.

[Kellnhofer *et al.*, 2019] 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.

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

[Li and Li, 2015] Jianfeng Li and Shigang Li. Gaze estimation from color image based on the eye model with known head pose. *IEEE Transactions on Human-Machine Systems*, 46(3):414–423, 2015.

[Li *et al.*, 2024] Guofa Li, Guanglei Wang, Zizheng Guo, Qing Liu, Xiyuan Luo, Bangwei Yuan, Mingrui Li, and Lu Yang. Domain adaptive driver distraction detection based on partial feature alignment and confusion-minimized classification. *IEEE Transactions on Intelligent Transportation Systems*, 25(9):11227–11240, 2024.

[Liu *et al.*, 2021] 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.

[Mazzamuto *et al.*, 2025] Michele Mazzamuto, Antonino Furnari, Yoichi Sato, and Giovanni Maria Farinella. Gazing into missteps: Leveraging eye-gaze for unsupervised

mistake detection in egocentric videos of skilled human activities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8310–8320, 2025.

[Muhammad *et al.*, 2020] Khan Muhammad, Amin Ullah, Jaime Lloret, Javier Del Ser, and Victor Hugo C De Albuquerque. Deep learning for safe autonomous driving: Current challenges and future directions. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4316–4336, 2020.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021.

[Rangesh *et al.*, 2020] Akshay Rangesh, Bowen Zhang, and Mohan M Trivedi. Driver gaze estimation in the real world: Overcoming the eyeglass challenge. In *2020 IEEE Intelligent vehicles symposium (IV)*, pages 1054–1059. IEEE, 2020.

[Ryan *et al.*, 2025] Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M. Rehg. Gaze-LLE: Gaze Target Estimation via Large-Scale Learned Encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28874–28884, 2025.

[Vuillecard and Odobez, 2025] 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*, pages 13508–13518, 2025.

[Wan *et al.*, 2021a] Zhong-Hua Wan, Cai-Hua Xiong, Wen-Bin Chen, and Han-Yuan Zhang. Robust and accurate pupil detection for head-mounted eye tracking. *Computers & Electrical Engineering*, 93:107193, 2021.

[Wan *et al.*, 2021b] Zhonghua Wan, Caihua Xiong, Wenbin Chen, Hanyuan Zhang, and Shiqian Wu. Pupil-contour-based gaze estimation with real pupil axes for head-mounted eye tracking. *IEEE Transactions on Industrial Informatics*, 18(6):3640–3650, 2021.

[Xiao *et al.*, 2025] Yunfeng Xiao, Xiaowei Bai, Baojun Chen, Hao Su, Hao He, Liang Xie, and Erwei Yin. De^{2}Gaze: Deformable and decoupled representation learning for 3d gaze estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3091–3100, 2025.

[Yang *et al.*, 2021] Yirong Yang, Chunsheng Liu, Faliang Chang, Yansha Lu, and Hui Liu. Driver gaze zone estimation via head pose fusion assisted supervision and eye region weighted encoding. *IEEE Transactions on Consumer Electronics*, 67(4):275–284, 2021.

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

[Yu *et al.*, 2025] Ting Yu, Yi Lin, Jun Yu, Zhenyu Lou, and Qiongjie Cui. Vision-Guided Action: Enhancing 3d human motion prediction with gaze-informed affordance in 3d scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12335–12346, 2025.

[Zhang *et al.*, 2015] 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*, pages 4511–4520, 2015.

[Zhang *et al.*, 2017] 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, 2017.

[Zhang *et al.*, 2020] 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*, pages 365–381, 2020.