
3DPE-Gaze:Unlocking the Potential of 3D Facial Priors for Generalized Gaze Estimation

Yangshi Ge Yiwei Bao Feng Lu*

State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University
{geyangshi, baoyiwei, lufeng}@buaa.edu.cn

Abstract

In recent years, face-based deep-learning gaze estimation methods have achieved significant advancements. However, while face images provide supplementary information beneficial for gaze inference, the substantial extraneous information they contain also increases the risk of overfitting during model training and compromises generalization capability. To alleviate this problem, we propose the 3DPE-Gaze framework, explicitly modeling 3D facial priors for feature decoupling and generalized gaze estimation. The 3DPE-Gaze framework consists of two core modules: the 3D Geometric Prior Module (3DGP) incorporating the FLAME model to parameterize facial structures and gaze-irrelevant facial appearances while extracting gaze features; the Semantic Concept Alignment Module (SCAM) separates gaze-related and unrelated concepts through CLIP-guided contrastive learning. Finally, the 3DPE-Gaze framework combines 3D facial landmark as prior for generalized gaze estimation. Experimental results show that 3DPE-Gaze outperforms existing state-of-the-art methods on four major cross-domain tasks, with particularly outstanding performance in challenging scenarios such as lighting variations, extreme head poses, and glasses occlusion.

1 Introduction

Gaze estimation has wide applications in computer vision, being an essential technology in scenarios such as human-computer interaction[13, 28, 29], virtual/augmented reality[25, 31], and driving monitoring[1, 19, 27]. Gaze estimation methods can be broadly categorized into two types: model-based approaches and appearance-based approaches. Model-based approaches calculate gaze direction by simulating the anatomical structure of the eyeball, offering higher accuracy but typically requiring specialized hardware and controlled environments. Appearance-based approaches, on the other hand, learn gaze mapping relationships directly from image features, offering broader application potential.

Appearance-based gaze estimation research has undergone a critical transition from focusing solely on eye regions to utilizing full-face information. Early research primarily extracted local features from eye regions[26, 20], but this approach overlooked the important contextual information provided by other facial areas. Zhang et al.[38] pioneered the use of full-face input and introduced a spatial weighting mechanism, significantly improving prediction accuracy. Since then, full-face input has become the mainstream approach in gaze estimation research.

However, the full-face input strategy is a double-edged sword: while it contains rich gaze-related structural information, the eye region occupies only a small proportion of the image, simultaneously introducing numerous domain-specific interferences (such as lighting variations, skin color differences, expression changes, etc.)[33]. These irrelevant information easily lead to models overfitting to

*Corresponding author.

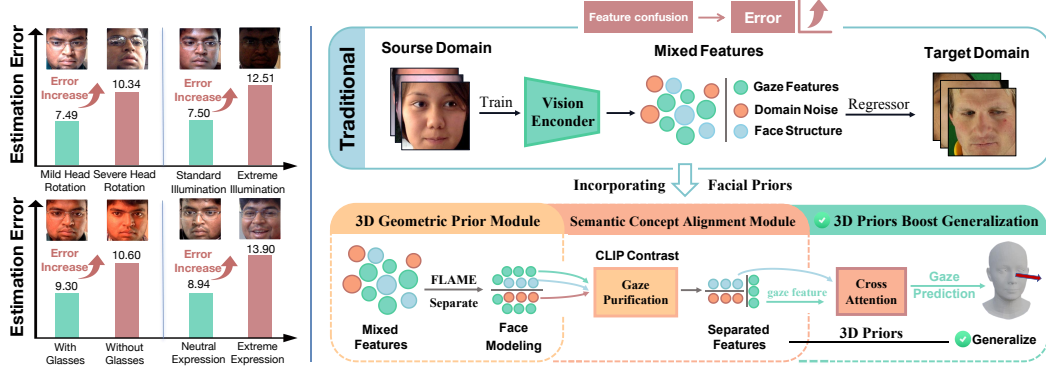


Figure 1: Left shows traditional CNN exhibiting clear performance collapse under standard conditions versus domain shift conditions. Right briefly demonstrates how we introduce 3D facial priors.

appearance features in the training data, severely affecting cross-domain generalization capability. To address this challenge, researchers have proposed various solutions, such as adversarial learning[7], data perturbation[33], feature separation[18], and multimodal fusion[36, 35]. However, most of these methods remain at the image feature level and fail to fully utilize 3D facial priors, making it difficult for them to capture the inherent geometric structure of the face and more susceptible to surface appearance changes. Therefore, **how to fully utilize structured facial information while effectively filtering domain-specific noise** remains a key challenge in the field of gaze estimation.

To address this challenge, we propose the Three-Dimensional Prior Enhanced Gaze Estimation framework (**3DPE-Gaze**), consisting of two complementary core modules: **3D Geometric Prior Module (3DGP)** and **Semantic Concept Alignment Module (SCAM)**. As shown in Figure 1, the 3DGP module decouples head pose, expression, shape, and gaze parameters from 2D facial images through the FLAME parametric face model, transforming gaze estimation from a prediction process based on domain-susceptible 2D appearance features to one based on stable geometric structures. However, FLAME-based encoders [9, 8] have inherent limitations in precisely modeling eyeball rotation and capturing subtle eye details, making it difficult to fully express complex gaze behaviors. To compensate for this deficiency, we designed the SCAM module as a complement. This module, based on CLIP semantic representations, distinguishes between gaze-relevant and irrelevant features through contrastive learning. Our FacePrior-Gaze Predictor effectively fuses these two complementary types of information, enabling the model to simultaneously leverage structured geometric constraints and high-level semantic concepts, thereby enhancing cross-domain generalization capability.

In summary, **3DPE-Gaze** systematically injects 3D facial priors and semantic concept priors into the gaze estimation process, effectively solving the cross-domain generalization problem in gaze prediction. Our experiments on multiple benchmark datasets have validated the excellent cross-domain performance of this method, achieving significant performance improvements in various challenging scenarios compared to current state-of-the-art methods. The main contributions of this research are as follows:

- We propose the 3D Geometric Prior Module (3DGP), a novel parametric encoder that transforms the gaze estimation problem from pixel space to parameters such as head pose, expression, and shape, establishing the foundation for associating gaze prediction with 3D facial structure and achieving effective extraction of domain-invariant features.
- We propose the Semantic Concept Alignment Module (SCAM), specifically addressing the limitations of pure geometric modeling in capturing complex gaze semantics. This module innovatively leverages CLIP pre-trained knowledge and contrastive learning strategies to explicitly distinguish gaze-relevant concepts from domain interference factors in the feature space, allowing the model to adapt to new environments without any target domain data.
- Experimental results show that our 3DPE-Gaze framework achieves significant improvements in various cross-domain settings. In cross-dataset evaluations, it reduces average error by more than 27% compared to the baseline and improves by 6% over SOTA methods, with particularly outstanding performance in highly challenging scenarios.

2 Related Work

Gaze Estimation. Gaze estimation has shifted from using local eye features [26, 20] to full-face inputs [38], which, despite including richer structural information, introduced domain-specific noise that impairs generalization. Existing solutions have largely focused on either purifying 2D image features [7, 33, 18] or introducing prior knowledge. For instance, some methods incorporate geometric cues like head pose as auxiliary information [40, 38, 22, 32, 17]. However, these approaches often treat geometric priors as supplementary inputs rather than core representations. In contrast, our 3DGP module fundamentally reframes the problem by using a parametric 3D model to map faces from an unstructured pixel space to a physically meaningful and decoupled parameter space. This transforms the task from simple "pixel-to-gaze" regression into a more robust "structure-to-gaze" prediction.

Other works have leveraged large-scale pre-trained models like CLIP to enhance robustness [35, 36]. These methods typically apply visual-linguistic knowledge to general 2D image features. Our work differs significantly in its design. The SCAM module performs a targeted semantic purification on a specific gaze code that has already been geometrically isolated by the 3DGP module. This novel "secondary purification" on a pre-decoupled feature is a key contribution of our framework.

Face Parametric Model. 3D face models have evolved from the 3DMM[2] first proposed by Blanz and Vetter, to BFM[23], FaceWarehouse[3], and then to FLAME[16], gradually enhancing the modeling capabilities for facial shape, expression, and movement. Although recent models like ICT[15] and FaceScape[34] offer higher geometric detail in appearance, they lack FLAME’s capabilities in feature disentanglement and efficiency. FLAME’s key advantage lies in its explicit parametric decoupling of shape, pose, and expression, a characteristic that highly aligns with the need in gaze estimation to separate gaze-related features from domain-specific interference. Additionally, it achieves a good balance between expressiveness and computational efficiency.

3 Method

3.1 3DPE-Gaze Framework Overview

To unlock the potential of 3D facial priors for generalized gaze estimation, we propose the **3DPE-Gaze** framework. As shown in Figure 2, the 3DPE-Gaze framework consists of two core modules and a final gaze predictor:

3D Geometric Prior Module (3DGP) The 3DGP module, based on the parametric face model FLAME, decomposes facial images into shape, expression, pose, lighting, and gaze parameters, achieving explicit decoupling of gaze features, facial structure features, and domain interference factors in physical space. This module also provides physically constrained anatomical foundations for gaze estimation through the reconstruction of 3D facial structures.

Semantic Concept Alignment Module (SCAM) The SCAM module utilizes CLIP pre-trained knowledge and contrastive learning strategies. This module explicitly distinguishes between “gaze-related concepts” (such as looking up-left) and “domain interference concepts” (such as lighting variations, facial expressions, etc.) in the feature space, providing cross-domain stable representations at the semantic level.

Facial Prior-Gaze Predictor The Facial Prior-Gaze Predictor integrates facial structural features extracted by 3DGP with semantic concepts identified by SCAM. This decoder focuses on highly gaze-relevant geometric regions through attention mechanisms while suppressing domain-specific interference such as background and lighting, generating the final accurate gaze prediction results.

This design successfully unlocks the potential of 3D facial priors in gaze estimation: the 3DGP module provides stable facial prior constraints, while the SCAM module compensates for FLAME’s limitations in eyeball modeling, offering semantic-level supplements. The complementary fusion of both significantly enhances the model’s cross-domain generalization capability. Technical details of each module will be elaborated in the following sections.

3.2 3D Geometric Prior Module (3DGP)

The 3DGP module decomposes input images into structured geometric parameter representations through a parametric face model, thereby transferring gaze estimation from the pixel space, which

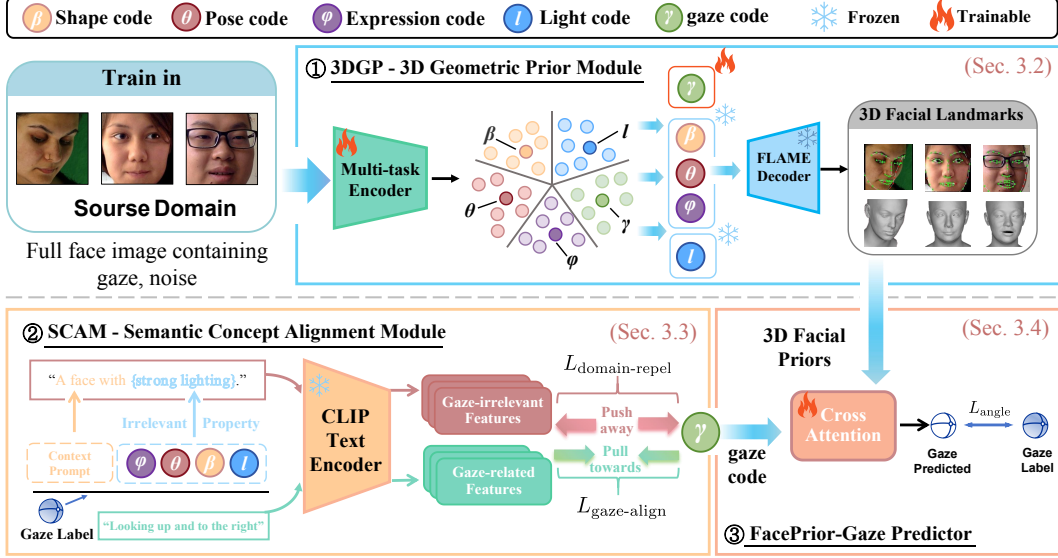


Figure 2: Overview of the proposed 3DPE-Gaze framework. The framework contains two core modules: (1) The 3D Geometric Prior Module (3DGP) utilizes the FLAME model to decompose facial images into geometric parameters, achieving physical decoupling of gaze features and domain interference; (2) The Semantic Concept Alignment Module (SCAM) separates gaze-related and unrelated concepts through CLIP-guided contrastive learning.

is susceptible to domain interference, to a more stable, decoupled parameter space. Through this decomposition, we can explicitly separate gaze code, facial structure parameters that affect gaze appearance, and domain-specific lighting parameters.

Parametric Representation and Encoding: Based on the FLAME model[16], we designed a gaze-specific multi-task encoder that receives preprocessed facial images I , and outputs a set of geometric parameters:

$$[\beta, \theta, \psi, \gamma, l] \quad (1)$$

where, $\beta \in \mathbb{R}^{100}$ controls static facial shape, $\theta \in \mathbb{R}^6$ represents head pose, $\psi \in \mathbb{R}^{50}$ captures dynamic expressions, $\gamma \in \mathbb{R}^{256}$ is a gaze-specific encoding containing eyeball movement information, and $l \in \mathbb{R}^{27}$ represents lighting parameters modeling environmental light conditions.

This parametric design decomposes facial images into domain-invariant geometric factors and susceptible appearance factors. For example, head pose θ and expression parameters ψ remain relatively consistent under different lighting or skin color conditions, while lighting parameters l explicitly capture environmental variation information. This explicit decoupling significantly reduces the model’s dependence on domain-specific appearance features.

To provide stable and structurally constrained facial parametric representations, we reused the encoder architecture of the DECA[9] pre-trained model (based on a ResNet backbone) and initialized all layers corresponding to the β, θ, ψ, l parameters in the multi-task encoder with its pre-trained weights (including the shared ResNet backbone and corresponding output branches). During training, these DECA-initialized parts are completely frozen (non-trainable) to ensure the accuracy and stability of facial structure parameters. In contrast, only the encoding branch responsible for outputting the gaze code γ is trainable. Considering that the γ parameter is a specific mapping of generic features extracted by the ResNet backbone to gaze direction, we designed this branch as a multi-layer perceptron (MLP) connected after the shared ResNet backbone output features, able to directly learn the mapping from robust features to gaze parameters, focusing on gaze feature learning.

3D Reconstruction and Structural Constraints: After obtaining FLAME parameters, we reconstruct the 3D facial mesh through the FLAME decoder:

$$M_{\text{pred}} = \text{FLAME}(\beta, \theta, \psi) = \bar{T} + B_S(\beta) + B_P(\theta) + B_E(\psi, \theta) \quad (2)$$

where $\bar{T}$ is the average face template, $B_S(\beta)$ is the shape blend shape controlled by identity parameters, $B_P(\theta)$ is the pose-related deformation, and $B_E(\psi, \theta)$ is the expression blend shape. From the reconstructed facial mesh, we extract key 3D facial landmarks $K \in \mathbb{R}^{N \times 3}$, which provide anatomy-based constraints for subsequent gaze decoding, ensuring that gaze predictions conform to physical laws.

Compared to traditional methods, the 3DGP module transforms gaze estimation from a black-box “pixel-to-gaze” mapping to a more structured “structure-to-gaze” prediction. By decoupling known geometric factors, it enhances model interpretability and significantly improves cross-domain stability.

3.3 Semantic Concept Alignment Module (SCAM)

The SCAM module compensates for 3DGP’s limitations in semantic understanding by explicitly distinguishing between “gaze-related” and “gaze-irrelevant” features at the concept level through the powerful vision-language alignment ability of the CLIP[24].

Systematic Generation of Semantic Concepts: To provide robust semantic supervision, we employ a systematic methodology for generating language descriptions for both gaze-related ($\mathcal{S}_{\text{gaze}}$) and domain interference ($\mathcal{S}_{\text{domain}}$) concepts.

For gaze-related concepts, a template-based approach is utilized, conditioned on the ground-truth gaze direction. We discretize gaze directions into canonical zones (e.g., “upper-right,” “lower-left”), from which textual prompts such as, “A person looking towards the upper-right,” are generated. This ensures the semantic representation is dynamically and accurately aligned with the precise directional gaze information for each sample.

For domain interference concepts, we leverage the physical parameters already extracted by our 3DGP module to programmatically generate a diverse set of gaze-irrelevant descriptions. This strategy creates a strong, coherent link between our geometric decoupling and semantic purification stages. Specifically, these prompts describe: **(1) Lighting conditions**, derived from the light parameters l (e.g., “under strong light,” “in dim light”); **(2) Facial expressions**, based on the expression parameters ψ (e.g., “a neutral expression,” “a person with a significant facial expression”); **(3) Facial shape**, informed by the shape parameters β ; and **(4) Head pose**, based on the pose parameters θ (e.g., “head tilted to the left”).

These systematically generated sets of descriptions are then passed through the CLIP text encoder to obtain their semantic representations:

$$t_{\text{gaze}} = \Psi_{\text{CLIP}}(s_{\text{gaze}}), \quad s_{\text{gaze}} \in \mathcal{S}_{\text{gaze}} \quad (3)$$

$$t_{\text{domain}} = \Psi_{\text{CLIP}}(s_{\text{domain}}), \quad s_{\text{domain}} \in \mathcal{S}_{\text{domain}} \quad (4)$$

where $\Psi_{\text{CLIP}}(\cdot)$ represents the CLIP encoder.

Contrastive Learning and Concept Separation: To align gaze code γ with gaze concepts in semantic space while keeping them away from domain interference concepts, we designed a bidirectional contrastive loss:

Gaze Concept Alignment Loss:

$$\mathcal{L}_{\text{gaze-align}} = -\log \frac{\exp(\text{sim}(z_\gamma, t_{\text{gaze}})/\tau)}{\sum_j \exp(\text{sim}(z_\gamma, t_j)/\tau)} \quad (5)$$

where z_γ is the projected gaze parameter, sim calculates cosine similarity, and τ is a temperature parameter.

Domain Interference Repulsion Loss:

$$\mathcal{L}_{\text{domain-repel}} = \frac{1}{|\mathcal{S}_{\text{domain}}|} \sum_{s \in \mathcal{S}_{\text{domain}}} \max(0, \text{sim}(z_\gamma, \Psi_{\text{CLIP}}(s)) - m) \quad (6)$$

where m is a boundary parameter, ensuring that gaze features maintain sufficient distance from domain interference concepts.

Through this contrastive mechanism, the SCAM module establishes an explicit boundary between gaze features and domain interference features in semantic space, enabling the model to focus on

high-level concepts related to gaze while filtering out domain-specific interferences. Complementary to geometric priors, semantic priors demonstrate unique advantages in handling abstract visual concepts and complex visual interferences.

3.4 Facial Prior-Gaze Predictor

To fully utilize both geometric and semantic priors, we designed a collaborative decoder that makes the two representations mutually enhancing through an attention mechanism, jointly guiding gaze prediction.

Cross-modal Feature Fusion: We adopt a cross-attention mechanism, using semantically guided gaze code γ as the Query, and three-dimensional geometric landmarks K as the Key and Value:

$$F_{\text{fused}} = \text{CrossAttention}(Q = W_q\gamma, ; K = W_kK, ; V = W_vK) \quad (7)$$

where W_q, W_k, W_v are learnable parameter matrices. This design enables the model to selectively focus on key structural regions based on semantic understanding, increasing sensitivity to gaze-related features.

Gaze Vector Prediction: The fused features generate gaze predictions through the final MLP decoder:

$$g = \frac{f_{\text{reg}}(F_{\text{fused}})}{|f_{\text{reg}}(F_{\text{fused}})|_2} \quad (8)$$

This collaborative mechanism achieves complementary enhancement of geometric and semantic priors: the structural features provided by 3DGP ensure that gaze predictions conform to physical constraints, while the semantic concepts contributed by SCAM guide the model to focus on the most relevant regions and suppress interference. Working together, they effectively alleviate the problem of traditional methods over-relying on domain-specific appearance features while fully utilizing structured information within the full-face range.

3.5 Training Objectives and Implementation

We adopt a multi-objective joint optimization strategy to train the 3DPE-Gaze framework, with the overall loss function:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{angle}} + \lambda_2 \mathcal{L}_{\text{gaze-align}} + \lambda_3 \mathcal{L}_{\text{domain-repel}} \quad (9)$$

Gaze Angular Loss $\mathcal{L}_{\text{angle}}$ measures the angular difference between predicted gaze and true gaze:

$$\mathcal{L}_{\text{angle}} = \arccos \left(\frac{\hat{g}^T g}{\|\hat{g}\|_2 \|g\|_2} \right) \quad (10)$$

Gaze Concept Alignment Loss $\mathcal{L}_{\text{gaze-align}}$ encourages gaze parameters γ to align with gaze-related concepts.

Domain Interference Repulsion Loss $\mathcal{L}_{\text{domain-repel}}$ ensures that gaze parameters γ stay away from interference concepts.

Hyperparameters λ_1, λ_2 , and λ_3 are used to balance the contributions of each loss term. Empirically, we set $\lambda_1 = \lambda_2 = \lambda_3 = 1.0$.

Training Implementation Details We conducted experiments on a single NVIDIA A100 GPU. Specifically, we adopted a staged training strategy: first freezing the FLAME parameter branches and pre-training the gaze parameter γ ; then introducing contrastive learning of the SCAM module and jointly optimizing the entire model. We used the Adam optimizer with an initial learning rate of 10^{-4} and a batch size of 256.

4 Experiments

Datasets and Evaluation Protocol. We adopt experimental settings consistent with cutting-edge research in cross-domain gaze estimation [7, 33, 36, 35, 18], evaluating our method on four

cross-domain tasks. Specifically, we use ETH-XGaze[39] and Gaze360[14] as training datasets, and MPIIFaceGaze[37] and EyeDiap[12] as testing datasets. For concise representation, we denote these four cross-domain tasks as $\mathcal{D}_E(\text{ETH-XGaze}) \rightarrow \mathcal{D}_M(\text{MPIIFaceGaze})$, $\mathcal{D}_E \rightarrow \mathcal{D}_D(\text{EyeDiap})$, $\mathcal{D}_G(\text{Gaze360}) \rightarrow \mathcal{D}_M$, and $\mathcal{D}_G \rightarrow \mathcal{D}_D$. This standardized cross-domain setup ensures that our experimental results can be directly compared with related research.

Data Preprocessing. For $\mathcal{D}_E$, $\mathcal{D}_M$, and $\mathcal{D}_D$, we normalize facial images following the standard method in [37]; for $\mathcal{D}_G$, we only select frontal face images to match the distribution characteristics of other datasets, which is consistent with former researches [7, 18]. All images are resized to a uniform resolution of 224×224 and normalized to the [0,1] range, thereby eliminating the influence of differences in acquisition devices and resolutions across different datasets.

4.1 Performance Comparison with State-of-the-Art Methods

Table 1: Performance comparison on cross-domain gaze estimation tasks (unit: degrees)

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$	Avg
CNN Baseline	8.56	8.90	9.51	8.48	8.86
PureGaze [7]	7.08	7.44	9.28	9.32	8.28
CDG [30]	6.73	7.95	7.03	7.27	7.25
Xu et al. [33]	6.50	7.44	7.55	9.03	7.63
Liang et al. [18]	5.79	6.96	7.06	7.99	6.95
CLIP-Gaze [36]	6.41	7.51	6.89	7.06	6.96
LG-Gaze [35]	6.45	7.22	6.83	6.86	6.84
Our 3DPE-Gaze	6.66	6.13	6.71	6.23	6.43

Table1 shows the performance comparison between 3DPE-Gaze and existing state-of-the-art methods on four cross-domain gaze estimation tasks. The results demonstrate that our method surpasses SOTA methods in 3 out of 4 cross-domain settings. From the overall performance, our approach exhibits stronger general generalization capability, achieving the lowest average error across all four cross-domain tasks, fully validating the excellent performance of our proposed framework in addressing cross-domain challenges. This result also confirms that effective utilization of facial priors can significantly enhance the cross-domain generalization capability of gaze estimation models.

Table 2: In-domain gaze estimation performance comparison (unit: degrees).

Method	within $\mathcal{D}_M$	within $\mathcal{D}_D$	within $\mathcal{D}_G$	within $\mathcal{D}_E$
Dilated-Net [4]	4.42	6.19	13.73	N/A
Gaze360 [14]	4.06	5.36	11.04	4.46
RT-Gene [11]	4.66	6.02	12.26	N/A
FullFace [38]	4.93	6.53	14.99	7.38
RCNN [21]	4.10	5.31	11.23	N/A
CA-Net [6]	4.27	5.27	11.20	N/A
GazeTR-Pure [5]	4.74	5.72	13.58	N/A
GazeTR-Hybrid [5]	4.00	5.17	10.62	N/A
CNN Baseline	4.74	7.49	13.23	5.69
Our 3DPE-Gaze	4.03	5.06	11.83	4.39

While our 3DPE-Gaze framework is primarily designed to enhance cross-domain generalization, it is also crucial to validate that this improvement does not compromise its performance within a single domain. Therefore, we conducted in-domain experiments, with the results presented in Table 2. The results show that our method achieves highly competitive, and in some cases state-of-the-art, performance. This demonstrates that our approach of leveraging 3D facial priors not only significantly boosts cross-domain robustness but also maintains excellent accuracy for in-domain gaze estimation.

4.2 Diagnostic Analysis of Geometric Prior Quality

To understand the performance discrepancy across different tasks, particularly the weaker result on the $\mathcal{D}_E \rightarrow \mathcal{D}_M$ task, we conducted a diagnostic analysis. We hypothesize that the model’s final accuracy is strongly correlated with the quality of the geometric priors extracted by the 3DGP module.

To test this, we use the stability of the reconstructed 3D Inter-ocular Distance (IOD) as a proxy for prior quality. For any given subject, their physical IOD is fixed; therefore, a lower standard deviation of the IOD across multiple images indicates a more stable and higher-quality 3D prior. As shown in Table 3, we grouped subjects from each target dataset into "High-Quality" and "Low-Quality" prior groups based on this metric.

The results clearly show that for both target domains, the High-Quality Prior group achieves significantly lower gaze error than the Low-Quality Prior group. This analysis confirms that our model’s performance is indeed dependent on the quality of the extracted 3D geometry, and the instability of priors from the $\mathcal{D}_M$ dataset is the primary reason for the higher error in that specific cross-domain task.

Table 3: Diagnostic analysis of geometric prior quality and its impact on gaze error. Avg. IOD Std Dev refers to the average standard deviation of the reconstructed 3D Inter-ocular Distance.

Domain	Prior Quality Group	Avg. IOD Std Dev	Average Gaze Error	Number of Subjects
$\mathcal{D}_M$	High-Quality Prior	4.1	5.50°	5
	Low-Quality Prior	10.5	7.24°	10
	Overall	8.4	6.66°	15
$\mathcal{D}_D$	High-Quality Prior	3.8	5.38°	9
	Low-Quality Prior	9.2	7.10°	7
	Overall	6.2	6.13°	16

4.3 Ablation Studies

Effectiveness of Core Modules

Our proposed framework relies on three complementary loss functions to achieve high-performance cross-domain gaze estimation: Gaze Angular Loss ($\mathcal{L}_{\text{angle}}$), Gaze Concept Alignment Loss ($\mathcal{L}_{\text{gaze-align}}$), and Domain Interference Repulsion Loss ($\mathcal{L}_{\text{domain-repel}}$). To verify the contribution of each loss function and their synergistic effect, we designed ablation experiments as shown in Table 4.

Table 4: Ablation experiments for different loss function combinations (unit: degrees)

Model Configuration	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
CNN Baseline	8.56	8.90	9.51	8.48
+ $\mathcal{L}_{\text{domain-repel}}$	7.32	7.19	7.63	7.44
+ $\mathcal{L}_{\text{gaze-align}}$	8.03	8.35	8.40	7.87
+ SCAM ($\mathcal{L}_{\text{gaze-align}} + \mathcal{L}_{\text{domain-repel}}$)	7.18	6.86	7.32	7.17
3DGP Only ($\mathcal{L}_{\text{angle}}$)	7.45	7.19	7.60	6.80
+ $\mathcal{L}_{\text{domain-repel}}$	6.90	6.50	7.05	6.35
+ $\mathcal{L}_{\text{gaze-align}}$	7.41	6.38	7.60	6.72
+ SCAM ($\mathcal{L}_{\text{gaze-align}} + \mathcal{L}_{\text{domain-repel}}$)	6.66	6.13	6.71	6.23

The results show that using the FLAME prior-based $\mathcal{L}_{\text{angle}}$ alone can significantly reduce errors, validating the fundamental value of geometric constraints. Further introducing $\mathcal{L}_{\text{gaze-align}}$ enhances model performance by leveraging semantic information. Finally, integrating $\mathcal{L}_{\text{domain-repel}}$ effectively separates gaze-related and unrelated features, bringing comprehensive performance improvements, with the complete model achieving the best average error. This indicates that the synergistic effect of geometric constraints, semantic guidance, and feature decoupling is crucial for improving cross-domain robustness.

Impact of Backbone Model Choices

To analyze the impact of our backbone model choices, we conducted ablation studies as suggested by reviewers, with results shown in Table 5. For the semantic encoder, we found that while more powerful CLIP architectures can improve performance, our chosen ViT-B-16 provides a strong balance between accuracy and efficiency. Similarly, for the FLAME parameter regressor, other high-quality models like SPECTRE also achieve competitive results, demonstrating our framework’s flexibility. We selected DECA as our primary regressor due to its wide availability and recognized strong performance.

Table 5: Ablation study on the impact of different backbone models (CLIP text encoders and FLAME parameter regressors). Unit: degrees.

Backbone Component	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
<i>Semantic Encoder (CLIP)</i>				
RN50	6.90	6.55	6.95	6.44
ViT-B-16 (Ours)	6.66	6.13	6.71	6.23
ViT-L-14	6.55	6.08	6.60	6.17
ViT-H-14	6.50	6.05	6.61	6.12
<i>FLAME Parameter Regressor</i>				
EMOCA [8]	6.85	6.40	6.90	6.45
DECA (Ours) [9]	6.66	6.13	6.71	6.23
SPECTRE [10]	6.59	6.10	6.65	6.18

Optimization of Geometric Feature Transfer Paths

Table 6: Ablation experiments for model architecture configurations (unit: degrees).

Model Configuration	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
Landmarks Only	6.66	6.13	6.71	6.23
Landmarks + Pose	7.20	6.70	6.85	6.47
Landmarks + Full Parameters	7.32	6.53	6.85	6.94

To identify the optimal geometric representation for generalization, we compared three feature configurations from the FLAME model: using only 3D landmarks, landmarks with head pose parameters, and landmarks with the full parameter set. As shown in Table 6, using only facial landmarks performed best across all tasks. This result suggests that landmarks provide a sufficiently structured abstraction of facial geometry, capturing key gaze-related relationships while filtering the domain-specific noise present in the more detailed parameters. Including the full parameter set introduced additional domain-specific biases that hindered performance, confirming that focusing on abstract, domain-invariant features is more effective for cross-domain learning.

4.4 Robustness Verification of Facial Priors in Extreme Scenarios

The core idea of our 3DPE-Gaze framework is to decouple gaze from other irrelevant features and incorporating 3D facial priors for generalized gaze estimation. Thus, in this section, we conduct experiments to verify the robustness of 3DPE-Gaze framework regarding varies factors, including head pose, lighting conditions, expression changes and glasses on the $\mathcal{D}_E \rightarrow \mathcal{D}_M$ task.

Robustness under Extreme Lighting Conditions. As shown in Figure 3 (left), our method outperforms the baseline under all lighting conditions, especially in extremes. It reduces errors by 8.3% in low-light and 10.6% in high-light areas, as our 3D geometric representation effectively separates facial structure from environmental lighting effects.

Adaptability to Expression Changes. Figure 3 (right) shows that our 3DPE-Gaze method consistently outperforms the baseline across the entire range of facial expressions. For extreme expressions (L1 intensity >17), our method maintains a stable error level while the baseline’s error significantly increases. This is due to our model’s ability to decouple expression muscle activity from gaze direction.

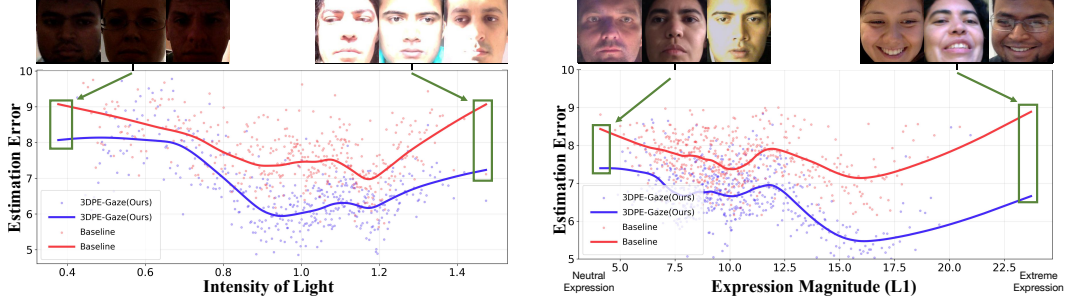


Figure 3: Robustness analysis under extreme scenarios. Left: Analysis of gaze estimation accuracy under different lighting intensities; Right: Impact of facial expression variations on gaze estimation precision.

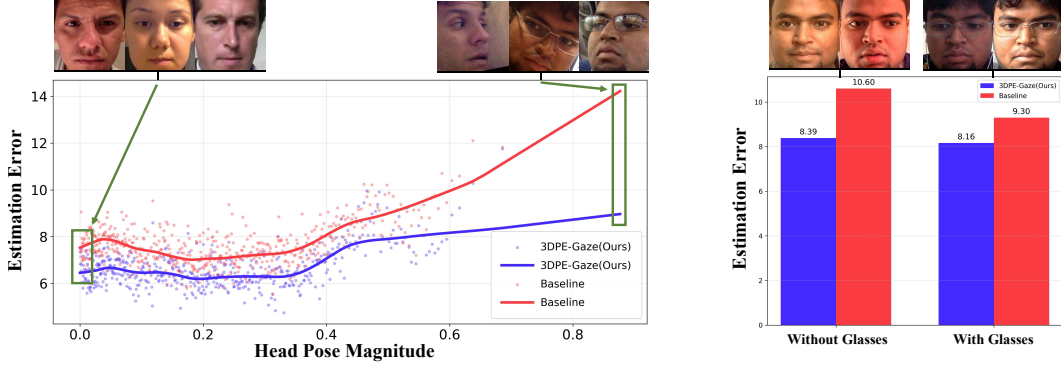


Figure 4: Robustness analysis under extreme scenarios. Left: Gaze estimation accuracy under large head rotations; Right: Impact of glasses occlusion on gaze estimation precision.

Large Head Rotations. Head pose variation is a major challenge in gaze estimation. As shown in Figure 4 (left), our method outperforms the baseline across all head poses, with the advantage growing at larger angles. By effectively separating the compound effects of head pose and eyeball movement, our method maintains stable performance even in extreme poses where the baseline fails.

Generalization Capability in Glasses Occlusion Scenarios. As shown in Figure 4 (right), our method significantly outperforms the baseline both without glasses (20.9% error reduction, $10.60^\circ \rightarrow 8.39^\circ$) and with glasses (12.3% error reduction, $9.30^\circ \rightarrow 8.16^\circ$). Most importantly, our model demonstrates excellent cross-condition stability, with an error difference of only 0.23° between the two conditions, compared to 1.30° for the baseline, ensuring a more reliable user experience.

5 Conclusion

This paper introduces 3DPE-Gaze, a novel framework for cross-domain gaze estimation that integrates 3D geometric and semantic priors. By leveraging a 3DGP module for geometric decoupling and a SCAM module for semantic purification, our method shifts the task from an unstable appearance-based space to a more robust geometric and semantic one. This design achieves state-of-the-art cross-domain performance without requiring any target domain data.

Limitations and Future Work. Limitations of our work include the FLAME model’s difficulty in precisely modeling fine eye details and our reliance on a predefined set of semantic concepts for contrastive learning. Future work will focus on three areas: developing specialized parametric eye models for finer detail; exploring adaptive semantic learning strategies, such as concept generation; extending the framework to dynamic, real-time applications on mobile and AR/VR devices.

Acknowledgement. This research has been supported by Beijing Natural Science Foundation (L242019).

References

- [1] Stefano Alletto, Andrea Palazzi, Francesco Solera, Simone Calderara, and Rita Cucchiara. Dr(eye)ve: A dataset for attention-based tasks with applications to autonomous and assisted driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.
- [2] Volker Blanz and Thomas Vetter. *A Morphable Model For The Synthesis Of 3D Faces*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. ISBN 9798400708978. URL <https://doi.org/10.1145/3596711.3596730>.
- [3] Chen Cao, Yanlin Weng, Shun Zhou, Yiyong Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3): 413–425, 2014. doi: 10.1109/TVCG.2013.249.
- [4] Zhaokang Chen and Bertram E Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer, 2018.
- [5] Yihua Cheng and Feng Lu. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3341–3347. IEEE, 2022.
- [6] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10623–10630, 2020.
- [7] 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.
- [8] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022.
- [9] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.
- [10] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos. *arXiv preprint arXiv:2207.11094*, 2022.
- [11] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European conference on computer vision (ECCV)*, pages 334–352, 2018.
- [12] 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, ETRA '14*, page 255–258, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327510. doi: 10.1145/2578153.2578190. URL <https://doi.org/10.1145/2578153.2578190>.
- [13] Christina Katsini, Yasmeen Abdrabou, George E. Raptis, Mohamed Khamis, and Florian Alt. The role of eye gaze in security and privacy applications: Survey and future hci research directions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20*, page 1–21, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376840. URL <https://doi.org/10.1145/3313831.3376840>.
- [14] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [15] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, et al. Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3410–3419, 2020.
- [16] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6), November 2017. ISSN 0730-0301. doi: 10.1145/3130800.3130813. URL <https://doi.org/10.1145/3130800.3130813>.

- [17] Yingxi Li, Xiaowei Bai, Liang Xie, Xiaodong Wang, Feng Lu, Feitian Zhang, Ye Yan, and Erwei Yin. Real-time gaze tracking via head-eye cues on head mounted devices. *IEEE Transactions on Mobile Computing*, 23(12):13292–13309, 2024. doi: 10.1109/TMC.2024.3425928.
- [18] Ziyang Liang, Yiwei Bao, and Feng Lu. De-confounded gaze estimation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 219–235, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73337-6.
- [19] Congcong Liu, Yuying Chen, Lei Tai, Haoyang Ye, Ming Liu, and Bert Shi. A gaze model improves autonomous driving. 06 2019. doi: 10.1145/3314111.3319846.
- [20] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. Adaptive linear regression for appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 36(10):2033–2046, 2014.
- [21] C Palmero, J Selva, MA Bagheri, and S Escalera. Recurrent cnn for 3d gaze estimation using appearance and shape cues. arxiv 2018. *arXiv preprint arXiv:1805.03064*.
- [22] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9368–9377, 2019.
- [23] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*, pages 296–301. Ieee, 2009.
- [24] 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. PmlR, 2021.
- [25] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutiérrez, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, PP, 12 2016. doi: 10.1109/TVCG.2018.2793599.
- [26] Kar-Han Tan, David J Kriegman, and Narendra Ahuja. Appearance-based eye gaze estimation. In *Sixth IEEE Workshop on Applications of Computer Vision, 2002.(WACV 2002). Proceedings.*, pages 191–195. IEEE, 2002.
- [27] Ashish Tawari, Kuo Chen, and Mohan Manubhai Trivedi. Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation. *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 988–994, 2014. URL <https://api.semanticscholar.org/CorpusID:19012458>.
- [28] Yunus Terzioğlu, Bilge Mutlu, and Erol Şahin. Designing social cues for collaborative robots: The role of gaze and breathing in human-robot collaboration. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, HRI '20*, page 343–357, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367462. doi: 10.1145/3319502.3374829. URL <https://doi.org/10.1145/3319502.3374829>.
- [29] Haofei Wang, Xujiong Dong, Zhaokang Chen, and Bert Shi. Hybrid gaze/eeg brain computer interface for robot arm control on a pick and place task. volume 2015, pages 1476–1479, 08 2015. doi: 10.1109/EMBC.2015.7318649.
- [30] 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.
- [31] Zhimin Wang and Feng Lu. Tasks reflected in the eyes: Egocentric gaze-aware visual task type recognition in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 30(11): 7277–7287, November 2024. ISSN 1077-2626. doi: 10.1109/TVCG.2024.3456164. URL <https://doi.org/10.1109/TVCG.2024.3456164>.
- [32] Yunfeng Xiao, Xiaowei Bai, Baojun Chen, Hao Su, Hao He, Liang Xie, and Erwei Yin. De2gaze: Deformable and decoupled representation learning for 3d gaze estimation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3091–3100, 2025. doi: 10.1109/CVPR52734.2025.00294.

- [33] Mingjie Xu, Haofei Wang, and Feng Lu. Learning a generalized gaze estimator from gaze-consistent feature. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i3.25406. URL <https://doi.org/10.1609/aaai.v37i3.25406>.
- [34] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 601–610, 2020.
- [35] 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. Springer, 2024.
- [36] Pengwei Yin, Guanzhong Zeng, Jingjing Wang, and Di Xie. Clip-gaze: towards general gaze estimation via visual-linguistic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6729–6737, 2024.
- [37] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2299–2308, 2017. doi: 10.1109/CVPRW.2017.284.
- [38] 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.
- [39] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V*, page 365–381, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58557-0. doi: 10.1007/978-3-030-58558-7_22. URL https://doi.org/10.1007/978-3-030-58558-7_22.
- [40] Wangjiang Zhu and Haoping Deng. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: In the abstract and introduction, we present previous work on cross-domain gaze estimation and claim that our work improves gaze estimation accuracy by introducing facial priors. We provide evidence for our algorithm in the experimental chapter, which reflects the claims made in the abstract and introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We have a subsection in the conclusion to discuss the limitations of our method and future research directions.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not contain theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: This paper fully discloses all the information needed to reproduce the main experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We have provided sufficient information to reproduce the results, and we will release the data and code upon cleanup.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have described all training details in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Due to limitations in computational resource access and training time, we were unable to calculate error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The computational resources required for the experiments are detailed in our paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We fully comply with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

There is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our models are not high risk and do not require safeguards.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cited the creators and complied with the licenses and terms of use.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Our paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.