Constrained Rotation Optimization: Revisiting Crop-Based Gaze Estimation

Riccardo Santambrogio[1], Jiawei Qin[2], Matteo Matteucci[1], and Yusuke Sugano[2]

[1] Politecnico di Milano, Milan, Italy
{riccardo.santambrogio, matteo.matteucci}@polimi.it
[2] The University of Tokyo, Tokyo, Japan
{jqin, sugano}@iis.u-tokyo.ac.jp

Abstract. Appearance-based gaze estimation typically relies on face normalization to reduce appearance variability, but this requires costly and error-prone landmark detection and head pose estimation. While crop-based alternatives have been explored, their geometric properties and performance trade-offs relative to normalization remain underexplored. In this work, we provide the first systematic comparison of normalization versus crop-based gaze estimation. To enable fair comparison, we formalize the crop-based approach through Constrained Rotation Optimization (CROp), making its geometric transformation explicit and comparable to normalization. We further adopt multi-task learning to recover head pose information lost in cropping. Through extensive experiments across various datasets, head pose distributions, and preprocessing conditions, we identify the conditions under which each approach excels. CROp shows advantages under extreme poses and noisy detection, while normalization benefits from landmark-based refinement in moderate conditions. Our analysis provides practical guidelines for choosing preprocessing strategies in real-world gaze estimation systems.

Keywords: Appearance-based gaze estimation

1 Introduction

Eye gaze is an important non-verbal cue that communicates information about human attention and intent. It plays a central role in numerous applications, such as human-computer interaction [48] and virtual/augmented reality [36]. For these reasons, achieving accurate gaze estimation is a relevant task in computer vision. Appearance-based gaze estimation methods have gained popularity in recent years due to their ability to map an eye or full-face image directly to their gaze direction in 3D space. Deep-learning models, such as convolutional neural networks (CNNs), can successfully learn this mapping from low-resolution images obtained using consumer-grade cameras, removing the need for specialized and expensive eye-tracking hardware.

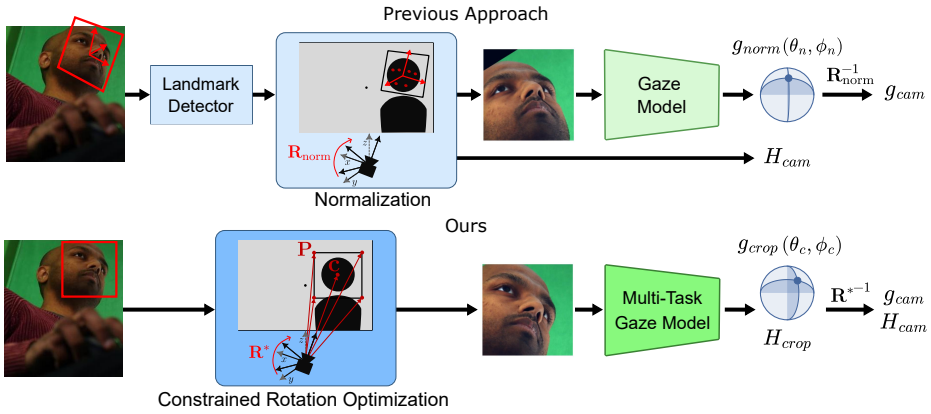


Fig. 1: Overview of our gaze estimation framework. Leveraging our Constrained Rotation Optimization mechanism, we estimate 3D gaze (pitch and yaw angles) directly from face crops, training a multi-task gaze model that simultaneously regresses head pose information.

Gaze estimation predicts a 3D vector in the camera coordinate system, but the high degree of freedom makes it difficult to learn the full range of appearance and gaze variations. To address this, a pre-processing step known as normalization or rectification [37, 47] is widely adopted. Normalization confines the image space and reduces appearance variability by standardizing head pose and scale. Since the normalized image corresponds to a virtual camera with a different orientation from the original, a rotation matrix is applied to convert gaze vectors between the two coordinate systems. Many prior works have been developed upon this task formulation [5, 8, 42, 43, 45].

However, normalization faces significant challenges: its performance depends on accurate facial landmark detection and head pose estimation, which can be computationally costly and are often unreliable in unconstrained environments. Alternative strategies that bypass explicit normalization have been explored. One category estimates gaze in specialized coordinate systems that model the relative positions of face and camera [23, 44], but the properties and limitations of these representations relative to normalization remain underexplored. Another method directly regresses the 3D gaze vector in the camera coordinate system from full images [4]. However, direct approaches introduce greater appearance variability, hindering generalization to novel environments.

In this work, we provide the first rigorous comparison between normalization and crop-based paradigms. A key observation is that face cropping, like normalization, changes the coordinate system in which the network operates. Like normalization, cropping shifts the principal point and thus changes the coordinate system in which the network predicts gaze. To formalize this, we introduce Constrained Rotation Optimization (CROp), which derives the coordinate change induced by cropping as a constrained rotation between the crop

and camera coordinate systems (Fig. 1). CROp establishes a mathematically precise mapping between the crop and the original camera coordinate systems, enabling accurate conversion of gaze vectors between them. This preserves the geometric consistency of normalization while removing its reliance on landmarks and explicit pose estimation. To compensate for the missing head pose input, we adopt a multi-task learning framework that jointly estimates gaze and head orientation using recent differentiable rotation representations [27, 34].

Our contributions are: (1) We make crop-based estimation comparable to normalization by formalizing its coordinate transformation (CROp), revealing its relationship to methods like Gaze360 [23] while providing mathematical rigor. (2) Through cross-dataset, within-dataset, and degradation experiments, we identify the conditions under which each approach excels. (3) We demonstrate that crop-based estimation, when properly formulated, achieves competitive or superior performance in realistic scenarios, offering practical deployment advantages.

2 Related Works

2.1 Appearance-Based Gaze Estimation

Appearance-based gaze estimation has emerged as a prominent approach due to its ability to operate under unconstrained conditions with minimal hardware demands [9,11,46,49]. Foundational work on data normalization [37,47] established essential strategies for handling variations in head pose and camera geometry. The introduction of large-scale datasets [10,26,45,51] has enabled the training of more robust and generalized models that maintain high accuracy across different environments. Recent efforts have continued to refine network architectures and training protocols [5, 7, 8, 33, 41–43, 45].

On the other hand, several works have explored gaze estimation without relying on explicit normalization. One approach is to directly map the face image to 2D coordinates on the screen [19,26]. However, these methods are constrained to fixed screen-based settings and do not generalize well to different camera viewpoints or arbitrary 3D gaze estimation. Other works explore an approach that outputs 3D gaze vectors from face images without normalization. Gaze360 [23] estimates gaze in an eye coordinate system defined from the center of the face. While their coordinate system shares similarities with our formulation, key differences exist: (1) they assume upright orientation (no roll), while CROp handles arbitrary rotations through optimization; (2) we explicitly derive the geometric transformation from camera coordinates given a crop, allowing systematic comparison with normalization; (3) we investigate the conditions under which crop-based estimation outperforms normalization, which was missing. GazeOnce [44] adopts a multi-task learning framework that jointly estimates gaze direction along with auxiliary tasks such as face localization. However, they do not explicitly estimate the 3D gaze vector in the camera coordinate system, making it difficult to determine the PoG in real-world scenarios. EFE [4] directly regresses the 3D gaze vector in the camera coordinate system using the entire camera image as input. While this avoids pre-processing, the wide appearance variability

increases task complexity, and its generalization performance in unseen environments remains uncertain. In contrast, we propose CROp that replaces the normalization process by modeling face cropping as a geometric transformation. This approach preserves the benefits of appearance constraints while eliminating the need for precise face alignment.

2.2 Head Pose Estimation

Head pose estimation is a widely studied task in computer vision due to its broad applications, including human-computer interaction and human behavior analysis. Most existing approaches focus on the 3 Degrees-of-Freedom (DoF) head rotation (yaw, pitch, and roll) [16, 18, 28], while some aim to estimate a 6 DoF pose that includes the translation [2, 3, 34].

Traditional methods have adopted the landmark-based approach [13,39] that localizes facial keypoints first, and subsequently aligns a 3D head model through Perspective-n-Point [20] algorithms. This approach, commonly used in gaze estimation pipelines, can produce highly accurate results in controlled environments but is sensitive to errors in landmark estimation [6]. To address these issues, research has shifted towards landmark-free approaches that directly regress head pose from images [1, 15, 35]. A key challenge for these methods is the choice of the rotation representation. Euler angles are intuitive but discontinuous and suffer from gimbal lock at large rotations [52].. Some works have used alternative representations such as quaternions [17] and 6D representations [15] to improve stability. Recently, a 9D continuous and differentiable representation named SVDO$^+$ was proposed [27] and first applied to head pose estimation in [34], demonstrating its effectiveness. We also adopt SVDO$^+$ to enable joint regression of gaze and 3 DoF head pose.

Similar to gaze estimation, when estimating head pose from a face image, it is not obvious how to revert it to the original camera coordinate system. Several works have explored predicting head pose from face crops while handling transformations to the original camera space. Img2Pose [2] formulates head pose estimation within an object detection framework, and directly regresses pose from crop proposals. IntrApose [34] further improves this by incorporating camera intrinsics. [28] defines a virtual camera rotation to rectify face images before performing head pose estimation. These methods highlight the importance of handling coordinate transformations when working with cropped images. Inspired by this, our approach extends the same idea to gaze estimation, ensuring accurate gaze recovery without requiring explicit normalization.

2.3 Gaze and Head Pose Joint Estimation

In addition to appearance-based approaches focusing solely on the eyes or face region, some methods explicitly integrate head and body pose information to handle more dynamic and unconstrained scenarios. [53] introduced a two-step approach that first extracts features via a CNN, then applies a parametric geometry-based model to compute 3D gaze angles. [30] proposed a method for

dynamic 3D gaze estimation from a distance by modeling temporal eye-head-body coordination, enabling robust performance in real-world settings. [25] proposed a weakly supervised approach with the look at each other (LAEO) loss, which depends on 3D head pose estimation. These works underline the significance of jointly leveraging head and body pose cues to enhance gaze estimation accuracy and robustness across diverse environments. Similarly, we propose a framework that jointly estimates gaze and head pose, addressing the challenge of unknown head rotation in cropped images without relying on facial landmarks.

3 Method

To enable fair comparison between normalization and crop-based gaze estimation, we must ensure both approaches have comparable geometric rigor. We address this by formalizing crop-based estimation through Constrained Rotation Optimization (CROp), which provides an explicit, mathematically grounded mapping between crop and camera coordinates.

3.1 Overview

Extracting a face crop from an image can be interpreted as defining a new *virtual camera* whose optical axis is centered on the crop. A gaze model predicts pitch and yaw relative to this optical axis, the same way normalization predicts them relative to the normalized camera axis. Because the crop axis points in a different direction from the original camera axis, the same 3D gaze direction corresponds to different pitch-yaw values in the two coordinate systems, so we need an explicit mapping that converts gaze between them. However, cropping is not equivalent to a rigid camera transformation, and no single rotation can exactly align the two views. We therefore formulate a constrained optimization problem that finds the best-approximating rotation by minimizing the projection error of a set of anchor points, namely the crop center and corners. The resulting rotation matrix provides a well-defined conversion between crop and camera coordinates. Importantly, this rotation is used only to convert gaze directions and does not synthesize the appearance of a rotated camera or modify the image pixels.

Unlike normalization pipelines, which explicitly estimate head pose during pre-processing, our crop-based formulation starts without head pose information. To complement this, we adopt a multi-task design where the model jointly predicts gaze and head orientation. This serves two purposes: it improves gaze accuracy by leveraging the natural coupling between head and eye movements, and it produces head pose estimates useful for downstream applications without additional pre-processing.

3.2 Constrained Rotation Optimization

To formalize our approach mathematically, we establish the relationship between the original camera coordinate system and that of the face crop. We define the

crop bounding box using the top-left (x_1, y_1) and bottom-right (x_2, y_2) corners. The camera intrinsic matrices of the original image $\boldsymbol{K}_{\text{cam}}$ and the one for the crop camera $\boldsymbol{K}_{\text{crop}}$ are defined as:

$$\boldsymbol{K}_{\text{cam}} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{K}_{\text{crop}} = \begin{bmatrix} f_x & 0 & (x_1+x_2)/2 \\ 0 & f_y & (y_1+y_2)/2 \\ 0 & 0 & 1 \end{bmatrix}. \tag{1}$$

Here, f_x and f_y are the focal lengths, (c_x, c_y) is the principal point in the full image, while the principal point in the cropped image coincides with the crop's center. Since the crop is a cut-out of the original image, its pixel coordinates differ only by translation. However, the different principal points mean that $\boldsymbol{K}_{\text{cam}}$ and $\boldsymbol{K}_{\text{crop}}$ define different optical axes: since the optical axis passes through the principal point, a shift in the principal point changes the 3D direction the axis points toward. Because gaze models predict yaw and pitch angles relative to this axis, the same gaze direction yields different angles in each system, and a rotation is needed to convert between them.

In the standard pinhole camera model, the camera matrix $\boldsymbol{K}_{\text{cam}}$ projects a point in 3D camera coordinates, $\boldsymbol{p}$, onto the image plane as $\boldsymbol{u} = [x_u \cdot s, y_u \cdot s, s]^\top = \boldsymbol{K}_{\text{cam}} \cdot \boldsymbol{p}$, where s is a non-zero scale factor. Conversely, given an image point expressed in homogeneous coordinates, $\boldsymbol{u} = [x_u, y_u, 1]^\top$, we can compute the ray intersecting it in camera coordinates (represented by a unit vector) by applying the inverse of the camera matrix $\boldsymbol{K}_{\text{cam}}$:

$$\hat{\boldsymbol{p}} = \frac{\boldsymbol{K}_{\text{cam}}^{-1}\boldsymbol{u}}{||\boldsymbol{K}_{\text{cam}}^{-1}\boldsymbol{u}||}. \tag{2}$$

The same operation applies to a point or vector in crop space, using $\boldsymbol{K}_{\text{crop}}$ instead of $\boldsymbol{K}_{\text{cam}}$.

We consider the center point of the bounding box $\boldsymbol{c} = [(x_1 + x_2)/2, (y_1 + y_2)/2, 1]^\top$, and the matrix consisting of stacked four corner points $\boldsymbol{P} \in \mathbb{R}^{3\times4}$ defined as

$$\boldsymbol{P} = \begin{bmatrix} x_1 & x_2 & x_1 & x_2 \\ y_1 & y_1 & y_2 & y_2 \\ 1 & 1 & 1 & 1 \end{bmatrix}. \tag{3}$$

Using Eq. (2), we compute the rays intersecting these points in camera coordinates, $\hat{\boldsymbol{c}}_{\text{cam}}, \hat{\boldsymbol{P}}_{\text{cam}}$, and the corresponding rays in crop coordinates, $\hat{\boldsymbol{c}}_{\text{crop}}, \hat{\boldsymbol{P}}_{\text{crop}}$ using $\boldsymbol{K}_{\text{crop}}$ instead of $\boldsymbol{K}_{\text{cam}}$. These rays represent the 3D directions from the respective camera origins through each point, providing the geometric relationship between the two coordinate systems.

Ideally, the desired coordinate transformation aligns the vectors in camera coordinates to the matching ones in crop coordinates, *i.e.*, a rotation $\boldsymbol{R}$ such that $[\hat{\boldsymbol{c}}_{\text{crop}}, \hat{\boldsymbol{P}}_{\text{crop}}] = \boldsymbol{R} \cdot [\hat{\boldsymbol{c}}_{\text{cam}}, \hat{\boldsymbol{P}}_{\text{cam}}]$. However, since a face crop is not a perfect rigid transformation in 3D space, this exact alignment is generally impossible to achieve. Therefore, we formulate an approximate solution.

We prioritize the alignment of the camera's optical axis (z-axis) to the center of the crop as our primary constraint, ensuring that the virtual crop camera is

directly facing the center of the face. Then, we seek the rotation $\boldsymbol{R}^*$ that optimally aligns the corner vectors, solving the constrained optimization problem:

$$\begin{aligned} \min_{\boldsymbol{R}\in SO(3)} \quad & \frac{1}{2}\|\hat{\boldsymbol{P}}_{\text{crop}} - \boldsymbol{R}\cdot\hat{\boldsymbol{P}}_{\text{cam}}\|_F^2 \\ \text{s.t.} \quad & \hat{\boldsymbol{c}}_{\text{crop}} = \boldsymbol{R}\cdot\hat{\boldsymbol{c}}_{\text{cam}}. \end{aligned} \tag{P}$$

In practice, $\boldsymbol{R}^*$ can be efficiently computed by finding the rotation that aligns $\hat{\boldsymbol{c}}_{\text{cam}}$ to $\hat{\boldsymbol{c}}_{\text{crop}}$ and then finding the rotation about $\hat{\boldsymbol{c}}_{\text{crop}}$ that optimally aligns the corner points. This can be solved using the Kabsch algorithm [21, 22, 38], which finds the optimal rotation between two sets of points. This operation is the core of our CROp approach.

To summarize, CROp computes $\boldsymbol{R}^*$ when given the crop bounding box and the camera intrinsic matrix $\boldsymbol{K}_{\text{cam}}$. Then, a gaze vector $\boldsymbol{g}_{\text{cam}}$ expressed in camera coordinates (*e.g.*, a label for a training sample) is converted to crop coordinates as $\boldsymbol{g}_{\text{crop}} = \boldsymbol{R}^*\cdot\boldsymbol{g}_{\text{cam}}$. A gaze direction predicted by the model in crop coordinates, parameterized by pitch and yaw angles (θ_c, ϕ_c), can then be converted back to camera coordinates as $\boldsymbol{g}_{\text{cam}} = \boldsymbol{R}^{*-1}\cdot\boldsymbol{g}_{\text{crop}}$. This is analogous to normalization, where $\boldsymbol{R}_{\text{norm}}$ plays the same role; the key difference is how the rotation is derived.

3.3 Multi-Task Gaze Model

Using CROp, we can estimate the gaze from simple face crops, avoiding the normalization procedure and thus ignoring head pose. To recover head pose within our pipeline, we adopt a multi-task model that learns it jointly with gaze. In detail, we repurpose our gaze model into one that predicts, from an input face image, both the gaze and the 3 DoF head pose (rotation). For gaze estimation, we denote the ground-truth gaze direction by $\boldsymbol{g}$ and the predicted gaze direction by $\hat{\boldsymbol{g}}$, and we use the angular loss:

$$\mathcal{L}_{\text{angular}} = \arccos\left(\frac{\hat{\boldsymbol{g}}\cdot\boldsymbol{g}}{\|\hat{\boldsymbol{g}}\|\|\boldsymbol{g}\|}\right). \tag{4}$$

For head pose, we adopt SVDO$^+$ [27], a 9D continuous and differentiable representation for rotations. This representation avoids the discontinuities and gimbal lock issues associated with Euler angles, making it particularly suitable for deep learning-based rotation regression. Calling $\hat{\boldsymbol{H}}$ the predicted head pose, we use the geodesic loss:

$$\mathcal{L}_{\text{geodesic}} = \arccos\left(\frac{\text{tr}(\boldsymbol{H}\hat{\boldsymbol{H}}^\top) - 1}{2}\right). \tag{5}$$

Our model is trained to optimize the multi-task loss

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{angular}} + \mathcal{L}_{\text{geodesic}}. \tag{6}$$

The head pose is predicted in the same crop coordinate system as the gaze and can be converted to camera coordinates using the transformation determined by CROp. Specifically, the head pose in camera coordinates is obtained

as $\hat{\boldsymbol{H}}_{\text{cam}} = \boldsymbol{R}^{*-1}\hat{\boldsymbol{H}}$. In addition to recovering head pose estimates, joint estimation is beneficial for gaze estimation accuracy, as shown in Sec. 4.3.

4 Experiments

We conduct systematic experiments in both cross-domain and within-domain settings to analyze under what conditions CROp outperforms normalization. We then ablate the components of our framework and demonstrate its robustness under strong and artificial degradation.

4.1 Experimental Settings

Gaze Datasets. We use four different gaze estimation datasets that provide the 3D eye gaze in camera coordinates, as well as accurate camera intrinsics and annotated head pose, to allow the comparison with normalization. **MPIIFaceGaze** [50] (MPII) contains 15 subjects looking at on-screen targets on their laptops under various lighting conditions, both indoor and outdoor. **ETH-XGaze** [45] (XG) is a large dataset collected using 18 high-resolution cameras in a studio environment. We use only the train set with publicly available annotations, comprising 80 subjects and 756K images. **EYEDIAP** [10] (ED) is a video dataset that contains 16 subjects gazing at continuous screen targets and 3D floating objects. **EVE** [31] is a video dataset containing 54 participants recorded while looking at on-screen stimuli. We extract frames from the videos with a stride of 10 for more efficient training.

Model Architecture. Both our CROp formulation and the proposed multi-task loss are model-agnostic and applicable to any backbone network for gaze estimation. We conduct our main experiments using the ResNet18 [14] backbone, which has been shown in prior work to be highly competitive with larger networks for gaze estimation [29, 44]. In Sec. 4.4 we further demonstrate the versatility of our approach across different architectures.

Data Pre-processing. For normalization (Norm), we follow the standard procedure with focal length 960, camera distance 300, and a 448×448 ROI to ensure consistency across datasets. Gaze estimation benchmarks typically apply normalization using facial landmarks or head pose annotations provided by each dataset. These annotations are obtained either through manual labeling [50] or through dataset-specific geometric procedures, such as fitting 3D morphable models to dense depth data [10] or leveraging person- and setup-specific calibration [31]. To cover a broader range of inference scenarios, we evaluate two preprocessing settings. In the dataset-annotation setting **(GT)**, Norm relies on the annotated landmarks or head pose, while CROp receives bounding boxes centered on the annotated facial landmarks, with size proportional to the maximum landmark distance (scaled by 1.5, or 2 in EYEDIAP, where only two eye landmarks are available). In the detector setting **(Det)**, both methods rely on a unified landmark pipeline based on InsightFace [12]. Norm uses the detected bounding boxes and 106 facial landmarks, followed by head pose estimation via

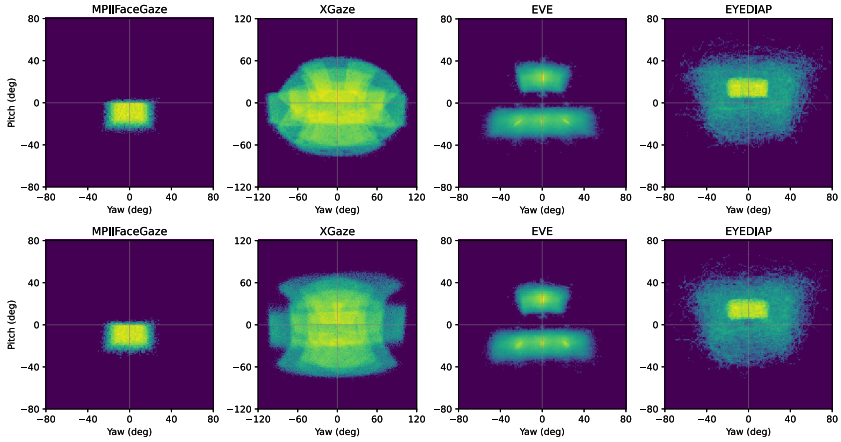


Fig. 2: Pitch-yaw gaze distributions induced by CROp (top) and normalization (bottom) across all datasets.

the PnP algorithm with a fixed generic head model. CROp instead uses only the detected bounding boxes and does not require landmarks.

Implementation Details. The inputs of gaze models are resized to 256 × 256. For CROp, the intrinsics matrix K_{crop} is computed from the original full-resolution image before resizing the input crop. During evaluation, gaze predictions estimated in crop coordinates are transformed back to the original camera coordinate system by inverting the $\boldsymbol{R}^*$ rotation before computing the angular error, as explained in Sec. 3.2. This ensures that all evaluation metrics are computed in the original camera coordinate system. Models are trained for 25 epochs using the Adam optimizer [24], with a learning rate of 1×10^{-4} decayed by 0.1 every 10 epochs.

4.2 CROp vs Normalization

We first evaluate whether our gaze estimation scheme, comprising CROp and joint head-pose learning, enables more accurate predictions than the standard normalization approach. The gaze distributions induced by the two approaches are visualized as pitch-yaw scatter plots across all datasets in Fig. 2, showing differences mainly at wide angles and outliers, with largely comparable overall angular ranges. In the following, we conduct cross-dataset and within-dataset evaluations, as well as error distribution analysis.

Cross-Dataset Evaluation We compare gaze models trained using the standard normalization scheme with our CROp approach in Table 1. Our approach achieves lower angular errors in most cases, demonstrating that gaze estimation can be effectively performed without normalization. Notably, the improvement is especially consistent on XGaze, where our method achieves an average error

Table 1: Cross-dataset evaluation of our approach (CROp and multi-task loss) versus normalization. We report results under both pre-processing from annotations (GT) and a face detector (Det). Values are angular errors (°).

Train	Method	Pre-proc.	MPII	XG	EVE	ED
MPII	Norm	GT	-	31.23	11.10	17.68
	Ours	GT	-	**29.37**	**10.19**	**15.46**
	Norm	Det	-	30.55	**11.43**	17.46
	Ours	Det	-	**29.56**	11.44	**15.59**
XG	Norm	GT	**7.19**	-	9.76	**12.15**
	Ours	GT	7.41	-	**8.49**	12.77
	Norm	Det	8.76	-	9.46	13.19
	Ours	Det	**7.76**	-	**8.98**	**12.70**
EVE	Norm	GT	9.69	39.71	-	21.32
	Ours	GT	**8.79**	**36.90**	-	**18.84**
	Norm	Det	9.53	39.13	-	20.47
	Ours	Det	**8.51**	**36.01**	-	**18.47**

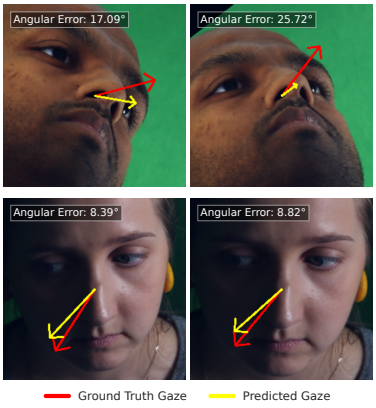


Fig. 3: Qualitative comparison of CROp using face crops (left) and normalization (right), on XGaze.

reduction of 6.52%, and on EYEDIAP, with a reduction of 10.61%, under ideal pre-processing (GT).

Our results in Tab. 1 also show that using a face detector (Det) does not significantly degrade accuracy in cross-dataset experiments. Indeed, a unified landmark detection pipeline can reduce systematic biases introduced by the distinct landmark estimation procedures adopted by different datasets. Under this setting, crop-based gaze estimation remains robust, achieving the best performance in virtually all cases when using automated face detection, which is most indicative of real-world deployment. This robustness is particularly evident on MPIIFaceGaze, where natural lighting and lower-quality laptop webcams can make normalization more susceptible to noise. These findings highlight the adaptability of our formulation, demonstrating reliable performance without the need for strict pre-processing procedures.

To understand the advantages of crop-based estimation, we analyze the error distribution in a particularly challenging case where it achieves strong performance: MPIIFaceGaze → XGaze. Although the average errors are large for both methods, our approach consistently outperforms Norm on a per-subject basis ($p = 6.65 \times 10^{-15}$, Wilcoxon signed-rank test [40]). Figure 4 shows how errors vary across head pose angles (pitch, yaw, and roll), computed in the original camera coordinate system, which is common to both methods. The crop-based representation tends to reduce errors across all head poses, but finds the largest relative improvements at extreme head angles. Although normalization is designed to mitigate head pose variation, we observe that it still struggles under extreme poses. Figure 3 illustrates this effect: for moderate head poses, our face-crop approach resembles normalized images, while at wider angles the two representations diverge significantly, but in both cases the resulting faces remain far from a canonical upright view. In these cases, direct estimation from face crops

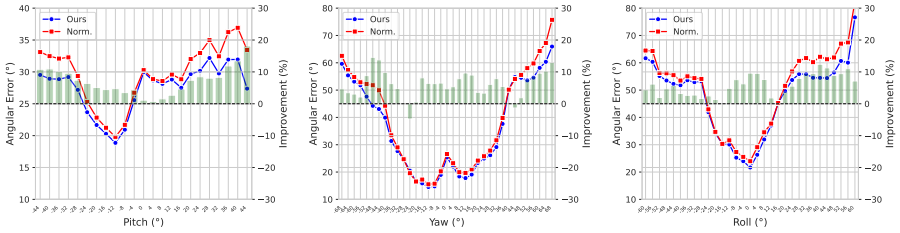


Fig. 4: Angular error as a function of head pose (pitch, yaw, roll) using CROp versus the normalization-based approach. The plots show absolute errors along with the relative improvement.

proves more reliable, suggesting that such representations can be advantageous in unconstrained conditions.

Within-Dataset Evaluation We further conduct within-dataset experiments on MPIIFaceGaze, XGaze, and EVE. For MPIIFaceGaze, we follow the standard 15-fold cross-validation protocol. To align with prior work, we use a 448×448 input size for this dataset. On XGaze, we train on the first 60 subjects and test on the remaining 20. For EVE, we report results on both the validation set (using publicly available annotations) and the test set (via the official server).

Unlike in cross-dataset settings, where multiple factors contribute to domain shift, within-dataset experiments isolate pre-processing noise as the primary source of variation. To mitigate its impact, we incorporate bounding box augmentations during training, applying rescaling and translations up to 20% of the bounding box size. Table 2 reports the results of this evaluation and shows that our approach consistently outperforms normalization, especially in the Det scenario, highlighting the robustness of crop-based gaze estimation under realistic conditions.

Computational Cost In Tab. 3 we show the runtime latency of our pipeline compared to normalization, breaking down the cost of each component (averaged over the XGaze dataset). Our pipeline avoids the landmark detector model entirely and saves 23% of the total computation time. Although CROp is formally expressed as an optimization problem, its implementation is efficient in practice and runs almost four times faster than normalization, which requires an image-warping operation.

4.3 Ablation Studies

In Tab. 4 we examine the two main components of our approach: the direct handling of crops using CROp, and the effectiveness of multi-task learning for gaze and head pose. To assess whether explicit head pose estimation benefits

Table 2: Within-dataset evaluation results of our approach (CROp and multi-task loss) versus normalization.

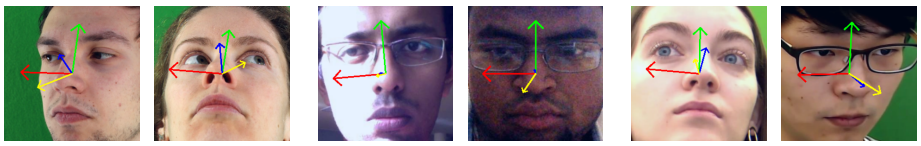
Method	Pre-proc.	MPII	XG	EVE Val	EVE Test
Norm.	GT	**4.93**	5.30	4.50	4.99
Ours	GT	5.02	**5.13**	**4.46**	**4.51**
Norm.	Det	5.32	6.45	4.71	5.10
Ours	Det	**4.98**	**5.55**	**4.69**	**4.99**

Table 3: Runtime comparison on RTX A6000. Numbers are ms.

Method	Face Det.	Lmk Det.	Norm.	CROp	Model	Total
Norm.	10.88	2.79	2.67	-	4.40	20.74
Ours	10.88	-	-	0.69	4.40	15.97

Table 4: Impact of head pose supervision. Models are tested under Det settings.

Train	Method	HPE	MPII	XG	EVE	ED
MPII	Norm.		-	30.55	11.43	17.46
	Norm.	✓	-	29.56	**11.30**	17.66
	CROp		-	**28.29**	12.18	16.72
	CROp	✓	-	29.56	11.44	**15.59**
XG	Norm.		8.76	-	9.46	13.19
	Norm.	✓	8.39	-	9.10	**10.55**
	CROp		8.05	-	9.99	12.26
	CROp	✓	**7.76**	-	**8.98**	12.70
EVE	Norm.		9.53	39.13	-	20.47
	Norm.	✓	9.10	37.03	-	20.76
	CROp		8.65	37.00	-	21.13
	CROp	✓	**8.51**	**36.01**	-	**18.47**



(a) MPIIFaceGaze →XGaze. **(b)** XGaze →MPIIFaceGaze **(c)** XGaze →EVE

Fig. 5: Qualitative results of CROp estimating gaze direction (yellow) and head pose (RGB axes).

gaze due to their inherent coupling, we also include results for a multi-task model trained on normalized face images.

Using CROp without joint learning of head pose already provides an advantage over standard normalization in all scenarios. Adding head pose supervision further improves overall performance, allowing the full multi-task formulation to achieve the best results in most cases. While a few cases show a slight reduction in gaze accuracy, this likely reflects the model's balancing between the two correlated objectives. However, this trade-off is offset by the benefit of obtaining head pose information directly, without additional computational steps. We also observe that multi-task loss benefits the normalization-based approach, although in this case, the head pose is already computed during pre-processing, making explicit estimation redundant. In contrast, introducing head pose estimation in the crop-based setting enables efficient recovery of head pose directly from the input image while still improving gaze estimation performance, without adding to the overall runtime (Fig. 5).

4.4 Additional Analyses

Bounding Box Noise and Occlusions We first demonstrated CROp 's adaptability to a real-world face analysis pipeline like InsightFace, without relying on manually curated bounding boxes. To further assess robustness, we introduce additional synthetic degradation on top of the detector's already realistic out-

Table 5: Impact of bounding box noise. Random scaling ($\pm X\%$) and translation ($X\%$ of box size); models trained on EVE and evaluated on MPIIFaceGaze.

Method	Bounding Box Noise					
	0%	10%	20%	30%	40%	50%
Norm.	9.53	9.53	9.53	**9.54**	**9.56**	**9.63**
Ours	**8.51**	**8.66**	**8.92**	**9.54**	10.25	10.83

Table 6: Impact of occlusions. Occlusions cover $X\%$ of bounding box area; models trained on EVE and evaluated on MPIIFaceGaze.

Method	Occlusion					
	0%	10%	20%	30%	40%	50%
Norm.	9.53	9.74	10.13	10.48	10.86	11.17
% Increase	-	2.2%	6.3%	10.0%	14.1%	17.2%
Ours	**8.51**	**8.68**	**8.93**	**9.22**	**9.58**	**9.90**
% Increase	-	2.0%	4.9%	8.3%	12.6%	16.3%

puts. Following the cross-dataset protocol, we take the models trained on EVE and evaluate them on MPIIFaceGaze, a dataset that is both visually challenging and sufficiently reliable in baseline performance. We introduce noise by applying random jittering and rescaling to bounding boxes, and simulate occlusions by masking face regions with randomly placed black boxes of varying sizes. These perturbations cover pessimistic conditions, with a broader range than what is typically encountered in practice.

As shown in Tab. 5, CROp remains stable under moderate bounding box noise, despite operating directly on noisy crops. Normalization benefits here from the additional step of facial landmark localization that can partially compensate for inaccuracies in the initial bounding box. Nevertheless, CROp outperforms normalization-based estimation up to $\pm 30\%$ translation and rescaling, indicating that explicit rectification is not required under realistic levels of bounding-box variation. At higher noise levels, normalization also starts to degrade, indicating failures in landmark detection. Table 6 presents results under increasing levels of occlusion. Because landmark detection is sensitive to occlusions, normalization suffers a significant performance drop as occlusion severity increases. Crop-based estimation is also affected, but degrades more smoothly and maintains lower errors across all occlusion levels. These results indicate that crop-based estimation provides reliable performance under realistic noise and occlusion, while normalization can compensate for bounding-box noise through additional preprocessing, but remains ultimately vulnerable to preprocessing failures.

Perspective Distortion and Extreme Poses A common challenge in gaze estimation is handling perspective distortion, especially when faces are close to the camera or observed at large off-axis angles. Normalization, which relies on a planar face assumption, is also affected under these conditions. We analyze this effect in Fig. 6 using the EVE dataset, which includes the widest viewing angles (up to $\sim 30°$) due to its left and right webcam views. While our method shows some degradation at large off-axis angles, where the approximation in $\boldsymbol{R}^*$ becomes less accurate, the same happens for normalization. Considering only the left and right cameras in EVE, and using models trained on XGaze, our method achieves a lower average error of $10.01°$ compared to $11.19°$ for normalization.

Table 7: Performance on synthetic XGaze. Numbers are angular errors (°).

Train	Norm.	Ours
MPIIFG	31.92	**30.74**
EVE	37.38	**34.34**

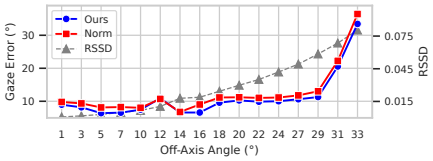


Fig. 6: Error vs off-axis angle on the EVE dataset. We report gaze-estimation error along with the CROp alignment residual, expressed as the root-sum-squared distance (RSSD); models trained on XGaze.

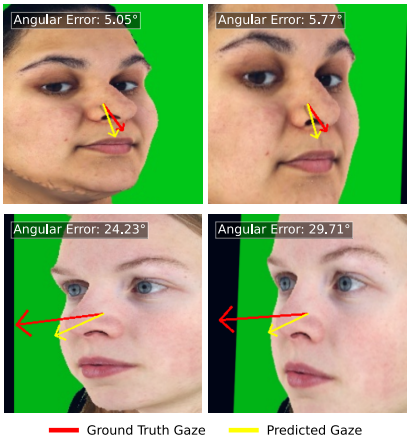


Fig. 7: Comparison of CROp (left) and normalization (right) on synthetic data with strong perspective distortion.

To further evaluate performance under more extreme conditions, we conduct experiments using synthetic data. Following previous work [32], we apply multi-view reconstruction to the ETH-XGaze dataset and synthesize images from arbitrary novel viewpoints. To obtain larger off-axis angles, we increase the camera's field of view, shift the face away from the optical axis, and closer to the camera. Specifically, we select the original frontal camera 0 and use the first five frames of each subject. For each selected image, we generate eight new views with varied camera positions, resulting in a total of 3,200 synthetic samples. This results in highly off-axis views of approximately 45°, introducing significant perspective distortion in the resulting face crops for both our method and normalization (see Fig. 7). As shown in Tab. 7, crop-based estimation continues to outperform normalization under these challenging synthetic conditions. Although gaze estimates are of limited practical use at such high error levels, these results confirm that estimation from face crops using CROp remains robust even under strong perspective distortion.

Different Backbones Our main experiments evaluated CROp using a ResNet18 backbone, but the proposed scheme is model-agnostic and compatible with any gaze estimation architecture. To assess this versatility, Tab. 8 reports results obtained with two additional popular backbones: ResNet50 and the transformer-based GazeTR-Hybrid [8] based on ResNet18. The results are consistent with our earlier findings, as CROp achieves the best performance in the majority of cases and remains highly competitive with normalization. Across backbones, the overall performance differences are modest, validating the suitability of ResNet18 for the main experiments, as larger models do not provide a substantial advan-

Table 8: Comparison of gaze estimation performance across different backbones (ResNet50 and GazeTR-Hybrid).

		ResNet50				**GazeTR-Hybrid**			
Train	**Method**	MPII	XG	EVE	ED	MPII	XG	EVE	ED
MPII	Norm.	-	34.45	14.47	19.75	-	29.46	12.15	19.94
	Ours	-	**30.77**	**11.52**	**17.14**	-	**27.00**	**11.21**	**12.57**
XG	Norm.	**6.30**	-	**7.19**	**9.21**	**7.20**	-	**9.42**	**10.29**
	Ours	7.06	-	8.00	11.75	8.44	-	9.87	14.58
EVE	Norm.	9.72	38.55	-	22.78	12.01	40.75	-	22.39
	Ours	**8.44**	**35.66**	-	**19.92**	**9.34**	**30.97**	-	**17.87**

tage. Normalization maintains a performance margin only on XGaze. This can be attributed to the challenging distribution of head poses in XGaze, which can steer the multi-task model toward head pose estimation during training. In the supplementary materials, we show that the multi-task model trained on XGaze learns accurate head pose predictions, reflecting the dataset's rich pose variation.

5 Conclusion

We presented the first systematic comparison of normalization versus crop-based preprocessing for gaze estimation. By formalizing crop-based estimation through CROp, we enabled a fair and controlled comparison across diverse conditions. Our analysis across datasets and degradation scenarios shows that CROp is most useful when landmark or head-pose preprocessing is unreliable or costly, with the largest gains under extreme poses and noisy detection. Normalization can retain an edge when accurate landmarks are available, under heavy synthetic bounding-box noise, or for some large-backbone models trained on XGaze, so its geometric precision may be preferable in controlled laboratory capture. However, for real-world deployment with consumer cameras and diverse poses, CROp provides a robust, efficient alternative. Our formalization bridges prior informal crop-based methods and rigorous normalization pipelines, enabling future work to make informed preprocessing choices based on application constraints.

While our method effectively estimates gaze direction, it assumes the gaze origin coincides with the bounding box center rather than the true anatomical origin point, and a rotation-based approximation of cropping. Future work could explicitly model and estimate the precise gaze origin within the facial structure, yielding more anatomically correct gaze vectors and improving overall accuracy.

Acknowledgements

This work was supported by the PNRR-PE-AI FAIR project funded by the NextGeneration EU program, the JST ASPIRE Program (Grant Number JPM-JAP2303), and the JST FOREST Program (Grant Number JPMJFR242R).

References

1. Ahn, B., Park, J., Kweon, I.S.: Real-time head orientation from a monocular camera using deep neural network. In: Proc. ACCV. pp. 82–96. Springer (2014)
2. Albiero, V., Chen, X., Yin, X., Pang, G., Hassner, T.: img2pose: Face alignment and detection via 6dof, face pose estimation. In: Proc. CVPR. pp. 7617–7627 (2021)
3. Algabri, R., Shin, H., Lee, S.: Real-time 6dof full-range markerless head pose estimation. Expert Systems with Applications **239**, 122293 (2024)
4. Balim, H., Park, S., Wang, X., Zhang, X., Hilliges, O.: Efe: End-to-end frame-to-gaze estimation. In: Proc. CVPRW. pp. 2688–2697 (2023)
5. Bao, Y., Lu, F.: From feature to gaze: A generalizable replacement of linear layer for gaze estimation. In: Proc. CVPR. pp. 1409–1418 (2024)
6. Chang, F.J., Tuan Tran, A., Hassner, T., Masi, I., Nevatia, R., Medioni, G.: Faceposenet: Making a case for landmark-free face alignment. In: Proc. ICCVW. pp. 1599–1608 (2017)
7. Cheng, Y., Bao, Y., Lu, F.: Puregaze: Purifying gaze feature for generalizable gaze estimation. In: Proc. AAAI. vol. 36, pp. 436–443 (2022)
8. Cheng, Y., Lu, F.: Gaze estimation using transformer. In: Proc. ICPR. pp. 3341–3347. IEEE (2022)
9. Cheng, Y., Wang, H., Bao, Y., Lu, F.: Appearance-based gaze estimation with deep learning: A review and benchmark. PAMI (2024)
10. Funes Mora, K.A., Monay, F., Odobez, J.M.: Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In: Proc. ETRA. pp. 255–258 (2014)
11. Ghosh, S., Dhall, A., Hayat, M., Knibbe, J., Ji, Q.: Automatic gaze analysis: A survey of deep learning based approaches. PAMI **46**(1), 61–84 (2023)
12. Guo, J., Deng, J.: Insightface: 2d and 3d face analysis project. `https://github.com/deepinsight/insightface` (2018)
13. Gupta, A., Thakkar, K., Gandhi, V., Narayanan, P.: Nose, eyes and ears: Head pose estimation by locating facial keypoints. In: ICASSP. pp. 1977–1981. IEEE (2019)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. CVPR. pp. 770–778 (2016)
15. Hempel, T., Abdelrahman, A.A., Al-Hamadi, A.: 6d rotation representation for unconstrained head pose estimation. In: ICIP. pp. 2496–2500. IEEE (2022)
16. Hempel, T., Abdelrahman, A.A., Al-Hamadi, A.: Toward robust and unconstrained full range of rotation head pose estimation. TIP **33**, 2377–2387 (2024)
17. Hsu, H.W., Wu, T.Y., Wan, S., Wong, W.H., Lee, C.Y.: Quatnet: Quaternion-based head pose estimation with multiregression loss. TMM **21**(4), 1035–1046 (2018)
18. Huang, B., Chen, R., Xu, W., Zhou, Q.: Improving head pose estimation using two-stage ensembles with top-k regression. Image and Vision Computing **93**, 103827 (2020)
19. Huang, Q., Veeraraghavan, A., Sabharwal, A.: Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. Machine Vision and Applications **28**, 445–461 (2017)
20. Huang, T.S., Bruckstein, A.M., Holt, R.J., Netravali, A.N.: Uniqueness of 3d pose under weak perspective: A geometrical proof. PAMI **17**(12), 1220–1221 (1995)
21. Kabsch, W.: A solution for the best rotation to relate two sets of vectors. Foundations of Crystallography **32**(5), 922–923 (1976)

22. Kabsch, W.: A discussion of the solution for the best rotation to relate two sets of vectors. Foundations of Crystallography **34**(5), 827–828 (1978)
23. Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., Torralba, A.: Gaze360: Physically unconstrained gaze estimation in the wild. In: Proc. ICCV. pp. 6912–6921 (2019)
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proc. ICLR (2015)
25. Kothari, R., De Mello, S., Iqbal, U., Byeon, W., Park, S., Kautz, J.: Weakly-supervised physically unconstrained gaze estimation. In: Proc. CVPR. pp. 9980–9989 (2021)
26. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: Proc. CVPR. pp. 2176–2184 (2016)
27. Levinson, J., Esteves, C., Chen, K., Snavely, N., Kanazawa, A., Rostamizadeh, A., Makadia, A.: An analysis of svd for deep rotation estimation. NIPS **33**, 22554–22565 (2020)
28. Li, X., Zhang, D., Li, M., Lee, D.J.: Accurate head pose estimation using image rectification and a lightweight convolutional neural network. TMM **25**, 2239–2251 (2022)
29. Liu, Y., Liu, R., Wang, H., Lu, F.: Generalizing gaze estimation with outlier-guided collaborative adaptation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3815–3824 (2021). https://doi.org/10.1109/ICCV48922.2021.00381
30. Nonaka, S., Nobuhara, S., Nishino, K.: Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In: Proc. CVPR. pp. 2192–2201 (June 2022)
31. Park, S., Aksan, E., Zhang, X., Hilliges, O.: Towards end-to-end video-based eye-tracking. In: Proc. ECCV. pp. 747–763. Springer (2020)
32. Qin, J., Shimoyama, T., Zhang, X., Sugano, Y.: Domain-adaptive full-face gaze estimation via novel-view-synthesis and feature disentanglement (2024), https://arxiv.org/abs/2305.16140
33. Qin, J., Zhang, X., Sugano, Y.: Unigaze: Towards universal gaze estimation via large-scale pre-training. arXiv preprint arXiv:2502.02307 (2025)
34. Roth, M., Gavrila, D.M.: intrapose: Monocular driver 6 dof head pose estimation leveraging camera intrinsics. IEEE Transactions on Intelligent Vehicles **8**(8), 4057–4068 (2023)
35. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. In: Proc. CVPRW. pp. 2074–2083 (2018)
36. Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., Wetzstein, G.: Saliency in vr: How do people explore virtual environments? IEEE transactions on visualization and computer graphics **24**(4), 1633–1642 (2018)
37. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3d gaze estimation. In: Proc. CVPR. pp. 1821–1828 (2014)
38. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. PAMI **13**(04), 376–380 (1991)
39. Werner, P., Saxen, F., Al-Hamadi, A.: Landmark based head pose estimation benchmark and method. In: Proc. ICIP. pp. 3909–3913. IEEE (2017)
40. Woolson, R.F.: Wilcoxon signed-rank test. Encyclopedia of biostatistics **8** (2005)
41. Xu, M., Wang, H., Lu, F.: Learning a generalized gaze estimator from gaze-consistent feature. In: Proc. AAAI. vol. 37, pp. 3027–3035 (2023)

42. Yin, P., Wang, J., Zeng, G., Xie, D., Zhu, J.: Lg-gaze: Learning geometry-aware continuous prompts for language-guided gaze estimation. In: Proc. ECCV. pp. 1–17. Springer (2024)
43. Yin, P., Zeng, G., Wang, J., Xie, D.: Clip-gaze: Towards general gaze estimation via visual-linguistic model. In: Proc. AAAI. vol. 38, pp. 6729–6737 (2024)
44. Zhang, M., Liu, Y., Lu, F.: Gazeonce: Real-time multi-person gaze estimation. In: Proc. CVPR. pp. 1–10 (2022)
45. Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., Hilliges, O.: Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In: Proc. ECCV. pp. 365–381. Springer (2020)
46. Zhang, X., Park, S., Maria Feit, A.: Eye gaze estimation and its applications. Artificial Intelligence for Human Computer Interaction: A Modern Approach pp. 99–130 (2021)
47. Zhang, X., Sugano, Y., Bulling, A.: Revisiting data normalization for appearance-based gaze estimation. In: Proc. ETRA. pp. 1–9 (2018)
48. Zhang, X., Sugano, Y., Bulling, A.: Evaluation of appearance-based methods and implications for gaze-based applications. In: CHI. pp. 1–13 (2019)
49. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proc. CVPR. pp. 4511–4520 (2015)
50. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: Proc. CVPRW. pp. 51–60 (2017)
51. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. PAMI **41**(1), 162–175 (2017)
52. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proc. CVPR. pp. 5745–5753 (2019)
53. Zhu, W., Deng, H.: Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In: Proc. ICCV. pp. 3143–3152 (2017)