

Article

Landmark-Guided Gaze Estimation via Conditional Keypoint Generation and Cross-Attention Fusion

Guanghai Xu ^{1,†}, Xiaoyang Zhang ^{1,2,†} , Wanli Zhao ^{2,†}, Zhongjie Mao ^{3,4}, Yue Li ^{2,5,*}, Duantengchuan Li ^{2,*}  and Liangshan Dong ⁶

¹ School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan 430068, China

² School of Information Management, Wuhan University, Wuhan 430072, China

³ School of Big Data and Artificial Intelligence, Chizhou University, Chizhou 247000, China

⁴ Wuhan HomeLightyear Technology Co., Ltd., Wuhan 430061, China

⁵ Department of Intelligent Construction, School of Civil Engineering, Wuhan University, Wuhan 430072, China

⁶ School of Physical Education, China University of Geosciences, Wuhan 430074, China

* Correspondence: yueli0301@whu.edu.cn (Y.L.); dtcleee1222@whu.edu.cn (D.L.)

† These authors contributed equally to this work.

Abstract

In gaze estimation, existing mainstream methods face significant challenges in capturing the fine-grained structures of eye regions, particularly in the absence of explicit geometric prior information, which hampers gaze prediction accuracy. To address this limitation, we propose the landmark-guided gaze estimation network (LGNet), a gaze estimation method guided by keypoints, which effectively incorporates geometric prior information to enhance estimation performance. The proposed method begins by training an eye-keypoint generator on the synthetic UnityEyes dataset using a Conditional Variational Autoencoder (CVAE). Next, we introduce a Symmetric Spatial Feature Fusion module (SSFF), combined with a dual-stream cross-attention mechanism, to achieve semantic alignment between the keypoint features and the facial image features extracted using ResNet50. Furthermore, we propose a Gated Channel Reweighting module (GCR) to suppress redundant information and amplify the critical features, thereby enhancing the model's overall response. Experimental results demonstrate that LGNet outperforms existing methods on three benchmark datasets. The code for this research has been made publicly available.

Keywords: gaze estimation; conditional variational autoencoder; deep learning



Academic Editor: Alessandra Lumini

Received: 12 December 2025

Revised: 30 January 2026

Accepted: 10 February 2026

Published: 25 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Eye movements serve as a direct reflection of human psychological processes, perceptions of the environment, and attentional focus on specific objects. Consequently, gaze estimation and eye tracking have emerged as prominent areas of research within artificial intelligence in recent years. Precise gaze estimation is critical for understanding user behavior and is instrumental in applications such as computer vision [1–4] and assisted driving [5]. Gaze estimation remains a highly complex and dynamic task. Achieving both high accuracy and the ability to adapt to a wide range of practical scenarios presents a significant challenge. This variability in gaze-related factors, coupled with the diverse application contexts, underscores the difficulty of developing a robust, versatile system for gaze estimation. Therefore, advancing gaze estimation methods that are both precise and adaptable remains a critical and ongoing research challenge.

Gaze estimation methods can primarily be categorized into two types: model-based approaches and appearance-based methods. Model-based methods rely on specialized hardware, such as infrared cameras or sensors implanted under the skin, to capture the intricate details of eye movements. In contrast, appearance-based methods utilize deep learning techniques to predict gaze direction from visual input, requiring large annotated datasets to achieve reliable results. The commonly used datasets in gaze estimation research are UnityEyes [6], MPIIGaze [7], MPIIFaceGaze [8], EyeDiap [9], Gaze360 [10], and ETH-XGaze [11]. However, the data collected in these datasets often come from a limited number of participants. Given the inherent structural differences in human eyes [12], there is a theoretical lower bound for gaze direction prediction accuracy in real-world applications. Additionally, factors such as lighting conditions, occlusions, and inconsistencies in image resolution can further degrade the accuracy of gaze estimation models. To mitigate these challenges, one promising direction is to employ binocular images in a complementary fashion for gaze direction regression, or to integrate additional semantic information to improve model accuracy [13,14]. Another effective strategy involves extracting keypoint information from critical regions of the eye to help guide the model's attention toward the most important areas. For instance, Seonwook et al. [15] introduced a gaze direction regression method based on the localization of eye keypoints, showing that keypoints can significantly enhance the personalization of gaze estimation tasks across multiple publicly available datasets. This approach has shown promise in improving gaze estimation accuracy, particularly in the face of varying conditions such as illumination and occlusion.

To overcome the challenges in gaze estimation, we propose LGNet, a 3D gaze estimation framework that integrates eye structural priors with facial image features to address challenges in gaze prediction. By utilizing eye keypoints, which contain rich geometric information correlated with gaze direction, our model improves accuracy and robustness to occlusions. Unlike traditional methods, we use CVAE to generate continuous, controllable keypoint predictions, enhancing the transferability of geometric priors across domains. Our method consists of a CVAE-based keypoint generator trained on synthetic UnityEyes data, which produces keypoint representations for real-world eye images. ResNet50 is used to extract facial features, which are fused with keypoints through a dual-stream cross-attention mechanism. To improve performance under low-quality images, we use keypoint coordinate vectors for supervision and incorporate an SSFF strategy for better spatial adaptability. Additionally, a GCR module filters redundant information, strengthening critical feature responses, while simultaneously reducing model parameters and inference overhead, thereby improving practical efficiency and deployability. Experiments show that LGNet achieves superior performance on several gaze estimation datasets, demonstrating strong robustness and accuracy in challenging conditions such as occlusion and low illumination. Our key contributions are summarized as follows:

- LGNet is presented as a landmark-guided gaze estimation framework that explicitly models eye geometry as a generative and transferable prior, fundamentally differing from recent transformer-based methods that rely on implicit structural learning from appearance features.
- A CVAE-based eye-keypoint generator is developed to produce continuous and controllable landmarks learned from synthetic data, which generalize effectively to real-world images and are actively utilized during inference as an independent geometric modality.
- To enable effective interaction between heterogeneous features, a dual-stream cross-attention fusion architecture is adopted, in which the proposed SSFF and GCR modules further enhance spatial alignment and suppress redundant channels, improving robustness under degraded visual conditions.

- Extensive experiments on three public benchmark datasets demonstrate that LGNet achieves superior accuracy and robustness, particularly under challenging conditions such as occlusion and low-quality inputs.

2. Related Work

Currently, appearance-based methods dominate the field of gaze estimation. These methods, which utilize deep learning techniques and do not require specialized hardware, have garnered significant research attention in recent years. The first Convolutional Neural Network (CNN)-based gaze estimation model, proposed by Zhang et al. [16], predicts gaze direction from a single-eye image by incorporating head pose information via a fully connected layer. However, Cheng et al. [13] highlighted the limitations of using single-eye input and proposed an asymmetric regression model that employs binocular images. Their model adaptively weights the two eyes to exploit complementary visual cues, improving gaze estimation accuracy. Further enhancing gaze estimation, Park et al. [14] introduced Gazemap, a semantic representation of eye images generated from gaze labels, which aids in single-eye gaze prediction.

To address robustness under challenging conditions such as low resolution and occlusion, Huang et al. [17] improved upon Gazemap by introducing a Hierarchical Conditional Variational Autoencoder (HCVAE) and a feature fusion module, reducing 3D gaze regression errors with just eye images. However, this approach does not incorporate head pose information, leading to overly complex models that require separate modules for each eye and head pose. In response, Zhang et al. [8] proposed a gaze direction prediction method based on full-face images. Their model used an auxiliary branch to emphasize the eye region, achieving high prediction accuracy with a simpler structure. To further refine gaze estimation, Li et al. [18] introduced ADGaze, which uses anisotropic Gaussian label distribution learning. Their approach, based on full-face input, improves robustness by using label distributions to guide coarse-to-fine classification and regression, showing superior performance on multiple public datasets. Some researchers have also incorporated eye-keypoint information into full-face-based gaze estimation to focus on the eye region. However, due to the lack of eye-region keypoint annotations in many gaze datasets, approaches often rely on synthetic datasets like UnityEyes and use semi-supervised or self-supervised techniques to detect keypoints. Gou et al. [19] proposed a cascaded Transformer framework that simultaneously detects eye keypoints, recognizes eye states, and performs gaze estimation, achieving improved performance through inter-task information sharing and the use of synthetic eye images for data augmentation.

Recent transformer-based and hybrid gaze estimation methods mainly rely on self- or cross-attention to model global dependencies from full-face images, sometimes incorporating auxiliary structural cues such as heatmaps or graph-based representations. However, in most cases, such structural information is either learned implicitly from appearance features or used only as auxiliary supervision during training. In contrast, LGNet explicitly models eye geometry as a generative latent representation. A Conditional Variational Autoencoder is employed to generate continuous eye keypoints, which serve as an explicit and transferable geometric prior and are actively utilized during inference rather than being limited to training-time supervision. Moreover, LGNet adopts a dual-stream cross-attention architecture to enable bidirectional interaction between appearance features and geometric priors, with the proposed SSFF and GCR modules further enhancing spatial alignment and channel reliability. This explicit geometry-guided design improves robustness under challenging conditions such as occlusion and low image quality.

3. Methods

The proposed LGNet is a gaze estimation framework that integrates keypoint and facial image features through cross-fusion. The architecture is illustrated in Figure 1. To address the challenge of limited keypoint annotations in gaze estimation datasets, we leverage the synthetic UnityEyes dataset [6] to train both a prior and a posterior network in a supervised manner. These networks share identical structures. The posterior network takes eye images and their corresponding keypoint labels as inputs, while the prior network is trained only with the eye images. This enables the prior network to predict keypoints for eye images in gaze estimation datasets during subsequent tasks. After training, two instances of the prior network are employed to generate keypoints for the left- and right-eye images separately. Next, multi-level features are extracted from the decoders of both prior networks. These features are combined using the SSFF module and fused with facial features extracted by a ResNet50 network. The fused features are then passed through a dual-stream cross-attention mechanism, which aligns the keypoints and facial features. Instead of directly concatenating the attention module's outputs, we introduce a GCR module to selectively retain the most informative features. Finally, the refined feature set is input into a regression head to produce the final 3D gaze direction prediction.

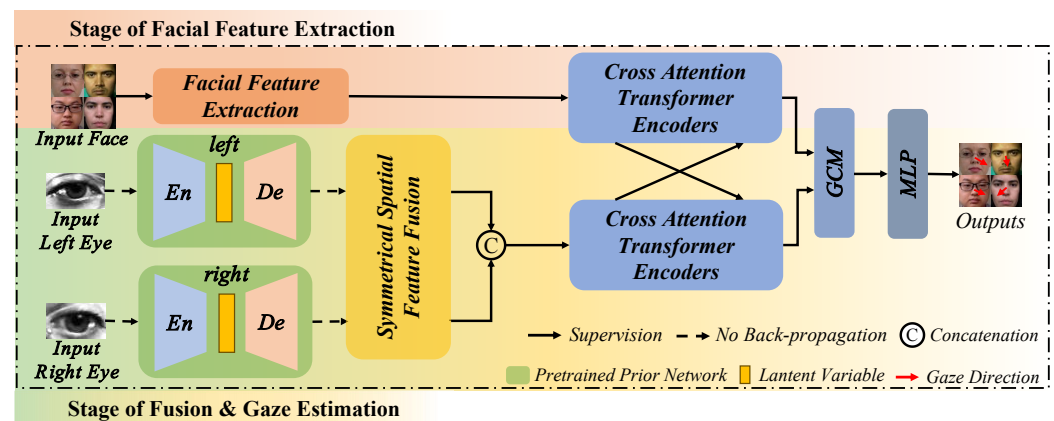


Figure 1. Framework of our LGNet model.

3.1. Keypoint Generation

We employ encoder–decoder networks based on the CVAE [20] framework to generate eye keypoints. First, we briefly analyze the Variational Autoencoder (VAE) [21] and explain how we apply CVAE in our task. As a generative model, the Variational Autoencoder (VAE) has the advantage of producing more plausible outputs. Given an input $\mathbf{x}$, the encoder $q_{\phi}(\mathbf{z}|\mathbf{x})$ of the VAE learns a distribution over latent variables, from which the decoder $p_{\phi}(\mathbf{x}|\mathbf{z})$ samples to reconstruct the input $\mathbf{x}'$. During inference, the objective of VAE is to maximize the marginal likelihood. For a given input sample $\mathbf{x}^{(i)}$, the log marginal likelihood is

$$\log p_{\theta}(\mathbf{x}^{(i)}) = D_{KL}(q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\theta}(\mathbf{z}|\mathbf{x}^{(i)})) + \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}). \quad (1)$$

In Equation (1), the KL divergence term involves the true posterior distribution $p_{\theta}(\mathbf{z}|\mathbf{x}^{(i)})$, which is intractable. Therefore, we introduce Equation (2):

$$\begin{aligned} \log p_{\theta}(\mathbf{x}^{(i)}) &\geq \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) = -D_{KL}(q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\theta}(\mathbf{z})) \\ &\quad + \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})} [\log p_{\theta}(\mathbf{x}^{(i)}|\mathbf{z})]. \end{aligned} \quad (2)$$

$\mathcal{L}$ denotes the variational lower bound. During VAE training, maximizing the evidence lower bound (ELBO) of the log-likelihood leads to model convergence. The first term of $\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})$ measures the divergence between two distributions. At the beginning of training, $\mathbf{z}$ is assumed to follow a standard Gaussian distribution. The second term measures the reconstruction error between the input and output of the VAE. Since the VAE decoder generates the output by sampling $\mathbf{z}$ from the latent space, the reparameterization trick $\mathbf{z} = g_\phi(\epsilon, \mathbf{x}) = \mu(\mathbf{x}) + \sigma(\mathbf{x}) \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is employed to enable backpropagation, where ϵ is a noise variable. In this work, the CVAE is utilized to perform the keypoint generation task. The CVAE takes both the image $\mathbf{x}$ (as conditional input) and the corresponding keypoint vector $\mathbf{y}$ as inputs and outputs the reconstructed keypoint vector $\mathbf{y}'$. Similar to the VAE, the ELBO for CVAE is defined as

$$\mathcal{L}_{\text{ELBO}}(\theta, \phi; \mathbf{x}, \mathbf{y}) = -KL(q_\phi(\mathbf{z} | \mathbf{x}, \mathbf{y}) || p_\theta(\mathbf{z} | \mathbf{x})) + \mathbb{E}_{q_\phi(\mathbf{z} | \mathbf{x}, \mathbf{y})} [\log p_\theta(\mathbf{y} | \mathbf{x}, \mathbf{z})]. \quad (3)$$

The approximate ELBO $\tilde{\mathcal{L}}_{\text{CVAE}}(\theta, \phi; \mathbf{x}, \mathbf{y}) \simeq \mathcal{L}_{\text{ELBO}}(\theta, \phi; \mathbf{x}, \mathbf{y})$ for each sample $\mathbf{x}^{(i)}$ can be written as

$$\tilde{\mathcal{L}}_{\text{ELBO}}(\theta, \phi; \mathbf{x}^{(i)}, \mathbf{y}^{(i)}) = -KL(q_\phi(\mathbf{z} | \mathbf{x}^{(i)}, \mathbf{y}^{(i)}) || p_\theta(\mathbf{z} | \mathbf{x}^{(i)})) + \frac{1}{L} \sum_{l=1}^L (\log p_\theta(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}, \mathbf{z}^{(i,l)})), \quad (4)$$

where $\mathbf{z}^{(i,l)} = g_\phi(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}, \epsilon^{(i,l)})$ is the latent variable sampled from the approximate posterior distribution $q_\phi(\mathbf{z} | \mathbf{x}^{(i)}, \mathbf{y}^{(i)})$, $\epsilon^{(i,l)} \sim \mathcal{N}(0, \mathbf{I})$, and l denotes the number of samples. In CVAE, compared to $p_\theta(\mathbf{z})$ in VAE, a conditional prior $p_\theta(\mathbf{z} | \mathbf{x})$ is used. This prior models the distribution $\mathbf{z}$ of given the input image $\mathbf{x}$, which can be arbitrary; here, we assume it to be Gaussian.

Hence, we introduce the model structure using the prior network as a representative. The prior network adopts a U-shaped structure similar to that of U-Net [22]. The encoder and decoder of the model each consist of four residual blocks; max-pooling or bilinear interpolation is applied between residual blocks. Additionally, downsampling in the encoder inevitably leads to information loss. To preserve the spatial details of the eyelid and eyeball in the eye image, each layer of En_{prior} uses a ResBlock with consistent input and output sizes. Each ResBlock uses 3×3 convolutional kernels with stride and padding set to 1. De_{prior} has connected with En_{prior} via skip connections. In addition, for the same reason as using ResBlocks, we incorporate a spatial attention mechanism (SA) on the skip connections.

During training, the input to the posterior network (referred to as the post network) consists of the synthetic dataset image $\mathbf{x}$ and the corresponding keypoint vector $\mathbf{y}$. The encoder En_{post} produces distribution parameters $\mathbf{z}_{\text{post}}(\boldsymbol{\mu}^{\text{post}}, \boldsymbol{\sigma}^{\text{post}})$ for the latent variable $\mathbf{z}$ associated with the keypoints, and the decoder De_{post} samples from this distribution to reconstruct the predicted keypoint vector $\hat{\mathbf{y}}_{\text{post}}$. The loss function of the post network is defined as follows:

$$\mathcal{L}_{\text{post}} = KL(q_\phi(\mathbf{z} | \mathbf{x}) || p_\theta(\mathbf{z})) + \|\mathbf{y} - \hat{\mathbf{y}}_{\text{post}}\|_2^2. \quad (5)$$

At the same time, $\mathbf{x}$ is also input into the prior network, whose encoder En_{prior} generates the distribution parameters $\mathbf{z}_{\text{prior}}(\boldsymbol{\mu}^{\text{prior}}, \boldsymbol{\sigma}^{\text{prior}})$ of the latent variable $\mathbf{z}$. Then, decoder De_{prior} samples from the distribution of $\mathbf{z}$ to produce the predicted keypoint vector $\hat{\mathbf{y}}_{\text{prior}}$. Since the prior network does not take the keypoint vector $\mathbf{y}$ as input, we employ KL

divergence to constrain the latent distribution of it, enforcing the latent distribution of the prior network to approximate that of the post network. This enables the prior to produce keypoint-related latent variables without relying on labeled keypoint input. Specifically, the KL divergence between the prior network and post network is defined as

$$\mathcal{L}_{KL} = \text{KL}(q_{\phi}(\mathbf{z} | \mathbf{x}, \mathbf{y}) \| p_{\theta}(\mathbf{z} | \mathbf{x}))$$

$$= \sum_{j=1}^J \left[\log \frac{\sigma_p^{(j)}}{\sigma_q^{(j)}} + \frac{(\sigma_q^{(j)})^2 + (\mu_q^{(j)} - \mu_p^{(j)})^2}{2(\sigma_p^{(j)})^2} - \frac{1}{2} \right], \quad (6)$$

where $\mu_q^{(j)}, \sigma_q^{(j)}$, and $\mu_p^{(j)}, \sigma_p^{(j)}$ denote the mean and standard deviation of the conditional prior and posterior Gaussian distributions for the j -th latent dimension, respectively. Both $q_{\phi}(\mathbf{z} | \mathbf{x}, \mathbf{y})$ and $p_{\theta}(\mathbf{z} | \mathbf{x})$ are modeled as multivariate Gaussian distributions with diagonal covariance matrices. The ELBO of the prior network is defined as

$$\mathcal{L}_{prior} = \mathcal{L}_{KL} + \|\mathbf{y} - \hat{\mathbf{y}}_{prior}\|_2^2. \quad (7)$$

The pretraining and inference procedures of the prior network are shown in Algorithm 1.

Algorithm 1 Generation of eye keypoints via conditional variational inference. During training, the posterior network is conditioned on both eye images and ground-truth keypoints from the UnityEyes dataset, while the prior network is conditioned only on eye images. During inference, the prior network alone is used to generate eye keypoints for eye images from the gaze estimation dataset

- 1: **Training:** Input eye image $\mathbf{x}_u$ and keypoints $\mathbf{y}_u$
- 2: Initialize parameters θ (posterior) and ϕ (prior)
- 3: **repeat**
- 4: **Posterior Encoder:** $(\mu^{post}, \sigma^{post}) \leftarrow \text{En}_{post}(\mathbf{x}_u, \mathbf{y}_u)$
- 5: Sample latent variable: $\mathbf{z}_{post} = \mu^{post} + \epsilon \odot \sigma^{post}$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
- 6: **Posterior Decoder:** $\tilde{\mathbf{y}}_u \leftarrow \text{De}_{post}(\mathbf{x}_u, \mathbf{z}_{post})$
- 7: **Prior Encoder (conditional on image only):** $(\mu^{prior}, \sigma^{prior}) \leftarrow \text{En}_{prior}(\mathbf{x}_u)$
- 8: Sample prior latent variable: $\mathbf{z}_{prior} = \mu^{prior} + \epsilon \odot \sigma^{prior}$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
- 9: **Prior Decoder:** $\hat{\mathbf{y}}_u \leftarrow \text{De}_{prior}(\mathbf{x}_u, \mathbf{z}_{prior})$
- 10: Compute loss(−ELBO):

$$\mathcal{L}_{post} = \text{KL}(q_{\phi}(\mathbf{z} | \mathbf{x}) \| p_{\theta}(\mathbf{z})) + \|\mathbf{y} - \hat{\mathbf{y}}_{post}\|_2^2.$$

$$\mathcal{L}_{prior} = \mathcal{L}_{rec}(\tilde{\mathbf{y}}_u, \mathbf{y}_u) + \text{KL}(q_{\phi}(\mathbf{z} | \mathbf{x}_u, \mathbf{y}_u) \| p_{\phi}(\mathbf{z} | \mathbf{x}_u))$$

- 11: Update (θ, ϕ) using the Adam optimizer (PyTorch 2.0.0).
 - 12: **until** convergence
-

- 13: **Inference:** Input eye images in gaze estimation datasets $\mathbf{x}_e$
 - 14: **Prior Encoder:** $(\mu^{prior}, \sigma^{prior}) \leftarrow \text{En}_{prior}(\mathbf{x}_e)$
 - 15: Sample latent variable: $\mathbf{z}_{prior} = \mu^{prior} + \epsilon \odot \sigma^{prior}$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
 - 16: **Prior Decoder:** $\hat{\mathbf{y}}_e \leftarrow \text{De}_{prior}(\mathbf{x}_e, \mathbf{z}_{prior})$
 - 17: **return** $\hat{\mathbf{y}}_e$
-

3.2. Keypoints' Feature Fusion and Gated Channel Reweighting

After pretraining of prior network, two prior networks were used to generate keypoints for the cropped left- and right-eye images, respectively. Since the keypoint feature fusion is performed on the decoder parts of the prior networks, the keypoints' feature fusion module in Figure 1 only shows the intermediate decoder layers of the two prior networks involved in feature fusion. Two sets of features, $F1$ and $F2$ ($F^* = \{F^0, F^1, F^2\}$), are extracted from the intermediate layers of the decoders of the two prior networks; these two features

are passed into the Symmetric Feature Fusion module to obtain fused keypoint features $F1_{kps}$ and $F2_{kps}$. Then, the two fused features are concatenated along the channel dimension to obtain F_{kps} . Next, the facial image features F_{face} extracted by ResNet50 and F_{kps} are fed into two Transformer encoders with dual-stream cross-attention mechanisms [23]. The outputs of the Transformer encoders are then passed through the proposed GCR module. Finally, a Multi-Layer Perceptron (MLP) head outputs the predicted 3D gaze direction.

3.2.1. Symmetrical Spatial Feature Fusion

To preserve spatial information as much as possible, we introduce spatial attention on top of the skip connections in the U-shaped keypoint generation network, and propose the SSFF module. Each prior network contains one Fusion Block, which allows keypoint-related features to provide reliable spatial guidance for subsequent cross-attention mechanisms, enabling robust implicit alignment between image features and keypoint features.

As shown in Figure 1, the Fusion Block is a multi-level feature fusion module, where semantically rich high-level features guide lower-level features in a coarse-to-fine manner. Specifically, we denote three decoder features from different scales as $R_1 \in \mathbb{R}^{B \times 256 \times H_1 \times W_1}$, $R_2 \in \mathbb{R}^{B \times 128 \times H_2 \times W_2}$, and $R_3 \in \mathbb{R}^{B \times 64 \times H_3 \times W_3}$, where B is the batch size.

In the first fusion stage, the highest-level feature R_1 is projected to the channel dimension of R_2 using a 1×1 convolution (Conv), followed by a transposed convolution (T-Conv) with kernel size of 4 and a stride of 2 to increase spatial resolution. Bilinear interpolation (BL) is then applied to ensure exact spatial alignment with R_2 . A channel-wise Softmax operation is used to generate a spatial attention map, which modulates R_2 via element-wise multiplication. The result is added back to R_2 through a residual connection to obtain the intermediate fused feature F_{middle} .

The second fusion stage follows the same procedure, where F_{middle} is first reduced in channel dimension via a 1×1 convolution, upsampled using transposed convolution and bilinear interpolation to match the resolution of R_3 , and then used to generate an attention map that guides R_3 . The final fused feature F_{final} is obtained via residual aggregation and serves as the output of the SSFF module.

$$\begin{aligned} F_{middle} &= R_2 + R_2 \odot \text{Softmax}_c \left(BL(T\text{-Conv}(\text{Conv}_{1 \times 1}(R_1))) \right), \\ F_{final} &= R_3 + R_3 \odot \text{Softmax}_c \left(BL(T\text{-Conv}(\text{Conv}_{1 \times 1}(F_{middle}))) \right). \end{aligned} \quad (8)$$

Here, R_1 , R_2 , and R_3 correspond to decoder features with channel dimensions 256, 128, and 64, respectively. Spatial resolutions are progressively increased through $T\text{-Conv}$ and BL to ensure exact alignment across scales.

3.2.2. Gated Channel Reweighting

In gaze estimation, different channels often carry significantly different semantic information. Especially under gaze deviation, occlusion, or illumination variations, certain channels may contain more critical region responses. Although cross-attention facilitates effective interaction between images and keypoint features, the original cross-fusion structure treats all channel features equally after fusion. Moreover, top-down multi-scale fusion may introduce redundant features. Without modeling the importance of each channel, key region responses may be diluted, reducing the overall regression accuracy. To address this, we propose a GCR module. It performs fine-grained channel-wise enhancement on the features after cross-attention. The architecture of this module is shown in Figure 1. It learns channel importance scores from global contextual information. A gating factor is introduced to adaptively control the weighting strength. This allows the model to better select and adapt features under varying sample conditions. Suppose the size of output

features from dual-stream cross-attention is $\mathbf{X} \in \mathbb{R}^{B \times N \times D}$; global average pooling is first applied to obtain a channel-wise context vector $\mathbf{z} \in \mathbb{R}^{B \times D}$. Then, an MLP is used to predict the channel importance vector $\mathbf{w} \in \mathbb{R}^{B \times D}$. To further adaptively adjust the weighting level, a global gating factor $\mathbf{g} \in \mathbb{R}^{B \times 1}$ is introduced. It controls the overall intensity of channel reweighting. The final output is denoted as $\mathbf{X}' \in \mathbb{R}^{B \times N \times D}$. The reweighting process is defined in Equations (9)–(12).

$$\mathbf{z} = \frac{1}{N} \sum_{i=1}^N \mathbf{X}_{:,i,:} \in \mathbb{R}^{B \times D}, \quad (9)$$

$$\mathbf{w} = \sigma \left(\mathbf{W}_2 \cdot \delta \left(\mathbf{W}_1 \cdot \mathbf{z}^\top \right) \right)^\top, \quad (10)$$

$$\mathbf{g} = \sigma \left(\mathbf{w}_g^\top \cdot \delta \left(\mathbf{W}_g \cdot \mathbf{z}^\top \right) \right), \quad (11)$$

$$\mathbf{X}'_{b,i,d} = \mathbf{g}_b \cdot \mathbf{w}_{b,d} \cdot \mathbf{X}_{b,i,d}, \forall b, i, d. \quad (12)$$

3.3. End-to-End Gaze Estimation

Once the 3D gaze prediction $\hat{\mathbf{g}}$ is obtained from our LGNet, we use the following loss function to compute the error:

$$\mathcal{L}_{\mathbf{g}} = \|\mathbf{g} - \hat{\mathbf{g}}\|_2^2, \quad (13)$$

where $\mathbf{g}$ is the ground truth of the 3D gaze vector. The total loss function for end-to-end training is defined as (where λ, η are regularization coefficients)

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{post} + \eta \mathcal{L}_{prior} + \mathcal{L}_{\mathbf{g}}. \quad (14)$$

4. Experiments

4.1. Datasets

To evaluate the performance of our proposed framework for appearance-based gaze estimation, we conducted experiments on three publicly available datasets: MPIIFaceGaze [8], EyeDiap [9], and Gaze360 [10]. These datasets are widely adopted benchmark datasets in gaze estimation, and using them enables a fair and consistent comparison with a substantial number of existing methods. The distributions of pitch and yaw angles across these datasets are depicted in Figure 2. For preprocessing, all three datasets undergo the same procedure: eye regions are cropped from normalized full-face images, and histogram equalization is applied to improve the contrast of the eye images.

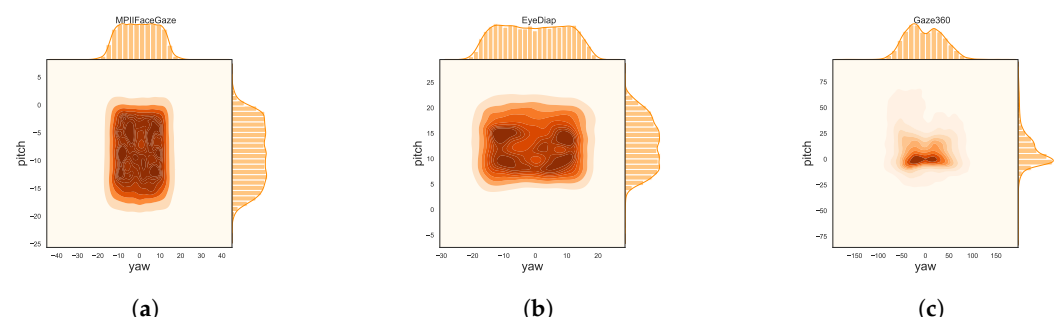


Figure 2. The distribution of yaw and pitch angles across the datasets is illustrated in Figure 2. Darker color regions indicate higher sample density, while lighter regions represent lower density. Specifically, panels (a,b) display the statistical distributions of gaze angles for the MPIIFaceGaze and EyeDiap datasets, respectively. It is important to note that due to preprocessing steps, which involve the removal of certain samples, the distribution shown in (c) does not fully represent the angular coverage of the original Gaze360 dataset.

- MPIIFaceGaze [8]: This dataset is widely recognized in appearance-based gaze estimation research. It contains data from 15 participants (9 male, 6 female, including 5 who wore glasses), recorded during various laptop usage scenarios where gaze targets randomly appeared on the screen. The dataset captures a range of environmental settings (indoor and outdoor), diverse lighting conditions, and various head poses. For model evaluation, we applied leave-one-subject-out cross-validation to assess performance on this dataset.
- EyeDiap [9]: This dataset comprises approximately 1000 video samples from 20 participants, with each video lasting 10 seconds at a resolution of 640×480 pixels. The videos exhibit variability in eye movement, head pose, and gaze direction over time. A distinctive feature of this dataset is its focus on dynamic scenarios, including rapid head movements, shifting gaze directions, and diverse background environments, making it ideal for testing the robustness and real-time performance of gaze estimation algorithms. We selected VGA-resolution videos containing screen-based targets, sampled one frame every 15 frames, and divided the participants into four folds for experiments using four-fold cross-validation.
- Gaze360 [10]: In this dataset, participants are instructed to focus on a moving target while their gaze directions are simultaneously recorded using multiple synchronized cameras. The dataset contains gaze data from 238 subjects (138 female, 100 male) collected across five indoor and two outdoor environments, resulting in a total of 172,000 samples. However, due to occlusions and gaze target positioning, the maximum pitch angle is limited to -40° and the yaw angle distribution is sparse, particularly near the extreme values. We evaluate the model's performance on the official test split provided by this dataset.

4.2. Training Details

For the keypoint generation network pretraining, we resize the input eye images to 90×150 pixels and apply various data augmentation techniques, including random flipping, salt-and-pepper noise, and random masking, to enhance the model's robustness. The training process consists of 100 epochs, with early stopping implemented using a patience value of 30 to prevent overfitting. During facial feature extraction, full-face images are resized to 224×224 pixels. To evaluate the model's performance, we employ leave-one-subject-out cross-validation on the MPIIFaceGaze dataset and four-fold cross-validation on the EyeDiap dataset, and use predefined training, validation, and test splits for the Gaze360 dataset. For optimization, we use the Adam optimizer [24] with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4} . All experiments are implemented in PyTorch (version 2.0.0) [25] and run on an NVIDIA A8000 GPU.

4.3. Evaluation Metrics

In the pretraining phase of the prior network, we evaluate performance using the Euclidean distance, as defined in Equation (15), which quantifies the difference between the predicted and ground-truth keypoint coordinates. Since the keypoint coordinates are normalized during preprocessing, the resulting Euclidean distance values fall within the range of $[0, 1]$. For the gaze estimation task, we employ angular error as the evaluation metric, calculated as outlined in Equation (16). This metric reflects the difference between the predicted and ground-truth gaze directions. The gaze labels are initially provided as 2D vectors representing yaw and pitch. These are first converted into 3D unit vectors, as described in Equations (17)–(19). The angular error is then computed using Equation (16). A lower angular error indicates a more accurate gaze prediction.

$$Dist = \left(\sum_{i=1}^{96} (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2 \right)^{\frac{1}{2}}, \quad (15)$$

$$Angle = \arccos \frac{\mathbf{g} \cdot \hat{\mathbf{g}}}{|\mathbf{g}| |\hat{\mathbf{g}}|}, \quad (16)$$

$$x = -\cos(pitch) \cdot \sin(yaw), \quad (17)$$

$$y = -\sin(pitch), \quad (18)$$

$$z = -\cos(pitch) \cdot \cos(yaw), \quad (19)$$

4.4. Performance Comparison

To evaluate the performance of the proposed LGNet framework for 3D gaze estimation, we conducted a thorough comparison with six state-of-the-art appearance-based gaze estimation methods. These comparisons were performed on three publicly available datasets: MPIIFaceGaze, EyeDiap, and Gaze360. The models chosen for this comparison include:

- (a) Dilated-Net: This is a lightweight gaze estimation model designed for mobile deployment, leveraging dilated convolutions to expand the receptive field. By using dilated kernels, the model retains high accuracy while significantly reducing the number of parameters, making it well-suited for resource-constrained devices.
- (b) Gaze360: This is a 3D gaze estimation approach tailored to unconstrained environments. It utilizes a panoramic camera to collect a large-scale dataset from 238 participants and employs a Long Short-Term Memory (LSTM) to integrate multi-frame temporal cues, which can provide additional robustness through temporal consistency.
- (c) CA-Net: This model incorporates a cross-attention mechanism to fuse information from both eyes. With a shared convolutional backbone, the left- and right-eye branches capture complementary cues, improving the overall gaze prediction accuracy by better integrating eye-specific features.
- (d) GTiT-Hybrid: This is a hybrid architecture that combines graph-based representations of local eye-region relationships with the global context captured by Transformers. This powerful fusion allows for enhanced gaze estimation accuracy, particularly in central-face gaze tasks.
- (e) GazeTR-Hybrid: This model combines a Transformer encoder with a CNN-based feature extractor, jointly encoding facial images and head pose into a unified representation. A Transformer decoder then generates the gaze direction, facilitating robust 3D gaze estimation in a variety of settings.
- (f) AGE-Net: This is a proxy-guided network that capitalizes on asymmetric features between the eyes. The network uses a backbone with two proxy regressors to extract eye-related and asymmetry cues, leveraging a joint loss function to improve both accuracy and stability in gaze prediction.

Table 1 presents a comparison of the average angular errors achieved by various methods across three datasets. The results show that LGNet outperforms all competing methods on MPIIFaceGaze and EyeDiap, achieving the lowest angular errors of 3.69° and 4.87°, respectively, which demonstrates the effectiveness of the proposed architecture. Specifically, on MPIIFaceGaze, LGNet achieves a noticeable improvement over GazeTR-Hybrid, highlighting the benefit of integrating eye keypoints with facial features for more accurate modeling of eye-region gaze information.

Table 1. Comparison of average angular errors (in degrees) achieved by different appearance-based gaze estimation methods on three public datasets: MPIIFaceGaze, EyeDiap, and Gaze360. Lower values indicate better performance. Leave-one-subject-out cross-validation is adopted for MPIIFaceGaze, four-fold cross-validation is used for EyeDiap, and the official train/validation/test split is applied for Gaze360. Results marked with * are reproduced by us under the same experimental settings for fair comparison. The best performance on each dataset is highlighted in bold.

Model	Input Type	MPIIFaceGaze	EyeDiap	Gaze360
Dilated-Net [26]	Face & Eyes	4.42	6.19	13.73
Gaze360 [10]	Full Face	4.06	5.36	11.04
CA-Net [27]	Two Eyes	4.27	5.27	11.20
GTiT-Hybrid [28]	Full Face	4.11	5.37	11.20
GazeTR-Hybrid [29]	Full Face	4.00	N/A	10.62
AGE-Net [30]	Two Eyes	3.83 *	4.96 *	12.3 *
LGNet (Ours)	Face & Eyes	3.69	4.87	11.03

On the Gaze360 dataset, LGNet achieves the second-best performance. This dataset is characterized by large head pose variations and highly unconstrained acquisition conditions, under which eye-region cues may become less reliable in extreme cases, partially limiting the effectiveness of geometry-guided eye modeling. Nevertheless, LGNet remains competitive and demonstrates good robustness under such challenging settings.

4.5. Ablation Study

4.5.1. Effectiveness of Each Module

To further evaluate the effectiveness of the proposed keypoints-guided mechanism and assess the contribution of each module to gaze estimation performance, we conducted a series of ablation studies on the MPIIFaceGaze, EyeDiap, and Gaze360 datasets. In these studies, we progressively incorporated key components and analyzed their impact on the overall performance. We began with a basic gaze estimation network as the baseline, which lacked the keypoint generation module. This baseline used ResNet50 to extract full-face features and directly regress the 3D gaze direction. Subsequently, we introduced the following components step by step:

- +CVAE—a conditional VAE-based prior network whose shallow decoder features are fused with the baseline.
- +CVAE+SSFF—the symmetric multi-scale feature fusion module is added on top of the keypoint features to enhance spatial guidance.
- +CVAE+SSFF+GCR—finally, a Gated Channel Reweighting (GCR) mechanism is added after the cross-attention output to explicitly model channel-wise importance and enhance feature representation.

Table 2 summarizes the results of the ablation study, clearly illustrating the contribution of each proposed module. Introducing the CVAE-based keypoint generation module consistently reduces the angular error across all datasets, confirming the effectiveness of geometric priors in guiding gaze estimation. The addition of the SSFF module further improves performance by enhancing multi-scale spatial alignment between keypoint features, which is particularly beneficial for datasets with diverse eye appearances and head poses. Finally, incorporating the GCR module yields the lowest errors on all datasets, demonstrating its ability to suppress redundant information and emphasize discriminative channel-wise features.

Overall, these results indicate that each component contributes from a complementary perspective, and their combination leads to a more accurate and robust gaze estimation framework.

Table 2. Ablation study of LGNet on MPIIFaceGaze, EyeDiap, and Gaze360 datasets. The table reports the mean angular error (in degrees) for different model configurations. The baseline model uses only full-face features without keypoint guidance. Each subsequent component is incrementally added to analyze its contribution to gaze estimation performance.

Model	MPIIFaceGaze	EyeDiap	Gaze360
Baseline	4.15	5.43	12.07
+CVAE	4.02	5.14	11.81
+CVAE+SSFF	3.88	4.91	11.36
+CVAE+SSFF+GCR	3.69	4.87	11.03

4.5.2. Pose-Stratified Analysis on Gaze360

To further analyze the effectiveness of the proposed landmark-guided prior under different pose conditions, we conduct a stratified evaluation on the Gaze360 dataset by grouping test samples according to yaw and pitch angles. Table 3 reports the mean angular error (MAE) of the baseline and LGNet across different angular ranges.

Table 3. Pose-stratified mean angular error (MAE) analysis on the Gaze360 dataset. Positive Δ indicates improvement over the baseline.

Type	Range (°)	Samples	Baseline MAE	LGNet MAE	Δ
Yaw	0–20	5217	8.146	7.614	0.533
Yaw	20–40	5611	7.996	7.425	0.571
Yaw	40–60	3530	9.402	9.432	−0.030
Yaw	>60	1673	18.000	19.087	−1.087
Pitch	0–20	11,877	6.221	6.036	0.184
Pitch	20–40	2639	8.808	9.060	−0.252
Pitch	40–60	943	12.298	11.022	1.276
Pitch	>60	572	17.453	15.367	2.086

As shown in the yaw-based analysis, LGNet achieves consistent improvements over the baseline for small to moderate yaw angles (0–20° and 20–40°). This indicates that when facial geometry and eye structures remain sufficiently visible, the landmark-guided prior can provide useful geometric cues to complement appearance features. However, for extreme yaw angles (>60°), LGNet underperforms the baseline. This degradation is likely caused by severe self-occlusion and unreliable keypoint estimation, which may introduce noisy or misleading geometric priors. In contrast, the pitch-based analysis reveals a different trend. LGNet demonstrates clear advantages under large pitch variations (40–60° and >60°), where the baseline error increases significantly. This suggests that when appearance-based cues become less stable due to eyelid deformation or illumination changes, the explicit geometric constraints introduced by landmarks help improve robustness and reduce angular error.

Overall, these results indicate that the contribution of the landmark-guided prior is condition-dependent: it is most effective under moderate yaw and large pitch variations, while its benefit becomes limited under extreme yaw angles with severe occlusion.

4.5.3. Computational Complexity Analysis

We further analyze the computational overhead introduced by the proposed GCR module. Table 4 reports the parameter counts, FLOPs, and inference latency of models with and without GCR, while keeping all other components (i.e., CVAE and SSFF) unchanged.

Table 4 reports the parameter count, FLOPs, and inference latency of the proposed GCR module compared with the baseline model without GCR. As shown, introducing GCR reduces the total number of parameters from 208.76 M to 137.98 M, achieving a 33.9% parameter reduction, while maintaining almost identical FLOPs and inference latency. This efficiency gain stems from the design of GCR, which explicitly reorganizes and filters redundant channel-wise features before the prediction head, leading to a more compact and informative feature representation. By suppressing less relevant channels and emphasizing discriminative ones, GCR effectively reduces the dimensionality of features fed into the subsequent prediction layers, thereby lowering parameter overhead without introducing additional computational burden.

Table 4. Comparison of model complexity and inference efficiency. Note that GCR reorganizes feature representations before the prediction head, resulting in a more compact input and reduced computational overhead.

Method	Params (M)	FLOPs (G)	Latency (ms)
w/o GCR	208.76	56.23	25.45
w/GCR (Ours)	137.98	56.16	25.37

Notably, the FLOPs (56.23 G vs. 56.16 G) and inference latency (25.45 ms vs. 25.37 ms) remain nearly unchanged, indicating that GCR incurs negligible runtime cost. These results demonstrate that GCR improves model compactness while preserving inference efficiency, highlighting its practical feasibility for real-world deployment, especially in scenarios where memory efficiency is critical.

4.6. Parameter Discussion

In LGNet, the overall training objective is formulated as a weighted combination of the gaze regression loss and the auxiliary keypoint-related losses. Consequently, the loss weights λ and η control the trade-off between enforcing geometric consistency and optimizing final gaze prediction accuracy. Understanding the sensitivity of model performance to these weights is therefore essential, especially for cross-dataset deployment. Instead of treating λ and η as dataset-specific hyperparameters requiring exhaustive tuning, we analyze the performance sensitivity trends induced by varying these weights. As shown in Figure 3, LGNet demonstrates a clear robustness to moderate variations of λ and η across all three datasets. Within a reasonably wide range, changes in loss weights result in only limited fluctuations in angular error, indicating that the proposed multi-term loss formulation is well-conditioned. Notably, although the optimal weight combinations differ slightly among datasets, consistent patterns can be observed. MPIIFaceGaze and EyeDiap, which share similar acquisition setups and head pose distributions, favor comparable weight configurations. In contrast, Gaze360—characterized by larger pose variations and more unconstrained environments—benefits from relatively smaller auxiliary loss weights. This behavior is expected, as the influence of geometric priors may diminish when extreme pose changes or partial occlusions reduce the reliability of eye-region cues. Importantly, these observations suggest that LGNet does not rely on precise, dataset-specific tuning of λ and η . Instead, the relative balance between gaze supervision and keypoint-related regularization plays a dominant role. This property is particularly desirable for cross-dataset generalization, as it allows the model to maintain stable performance when transferred to

new datasets with different characteristics, even when optimal hyperparameters are not explicitly re-optimized.

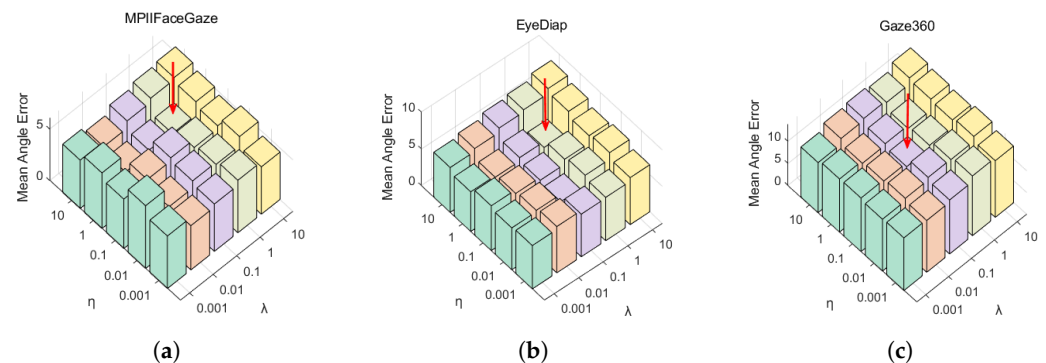


Figure 3. Mean angle error distributions of the LGNet model with different loss weight combinations (λ and η) on three datasets: (a) MPIIFaceGaze, (b) EyeDiap, and (c) Gaze360. The red arrows indicate the optimal parameter settings for each dataset. Lower bars correspond to better model performance.

Overall, the robustness of LGNet to loss weight variations highlights the stability of the proposed training objective and supports its practical applicability in real-world scenarios involving diverse data distributions.

4.7. Correlation Analysis

To assess the quality of the generated landmarks, we present several visualizations of keypoints. Specifically, we focus on subjects from the MPIIFaceGaze dataset, showcasing visualizations from the validation set. As illustrated, the pretrained keypoint generation network successfully produces high-quality landmarks on gaze estimation datasets. This performance is attributed to the dual advantages of using coordinate labels for supervision and integrating spatial attention modules, which collectively enhance the accuracy of keypoint generation and bolster spatial generalization capabilities. To further investigate the relationship between the predicted eye keypoints and gaze direction, we conducted a Spearman rank-correlation analysis between the keypoint coordinates and their corresponding gaze angles on the MPIIFaceGaze dataset, as shown in Figure 4. For this analysis, the right eye was selected as a representative example for single-eye analysis. This approach not only confirms the existence of a stable statistical relationship between keypoints and gaze direction but also highlights the significant role keypoints play in improving gaze estimation accuracy.

As illustrated in Figure 4a, pitch shows a significant correlation with the y-coordinate, indicating that the vertical displacement of the eyelid and iris plays a crucial role in vertical gaze shifts. This suggests that vertical gaze movements are primarily influenced by the up-and-down motion of the eye components. In contrast, yaw exhibits a stronger correlation with the x-coordinate, with noticeable peaks around keypoints 10–30. This pattern implies that horizontal gaze shifts are mainly driven by the horizontal rotation or displacement of the eyeball. To further explore this relationship, we performed a dual-eye analysis by combining the 48 keypoints from each eye into a single 96-dimensional landmark vector, and then computed the Spearman correlations between the x/y-coordinates and gaze components. As shown in Figure 4b, keypoints with a strong pitch–y correlation are distributed across both eyes, suggesting that vertical gaze shifts are perceived jointly by both eyes. In contrast, yaw correlations are more pronounced in specific right-eye keypoints, which implies that the network may place more emphasis on the right-eye region when estimating gaze direction. These findings demonstrate that the generated keypoints are spatially interpretable and exhibit robust, direction-specific statistical consistency with gaze direction.

This further validates the effectiveness of keypoint-based modeling for gaze estimation and highlights the contribution of keypoints in improving overall estimation accuracy.

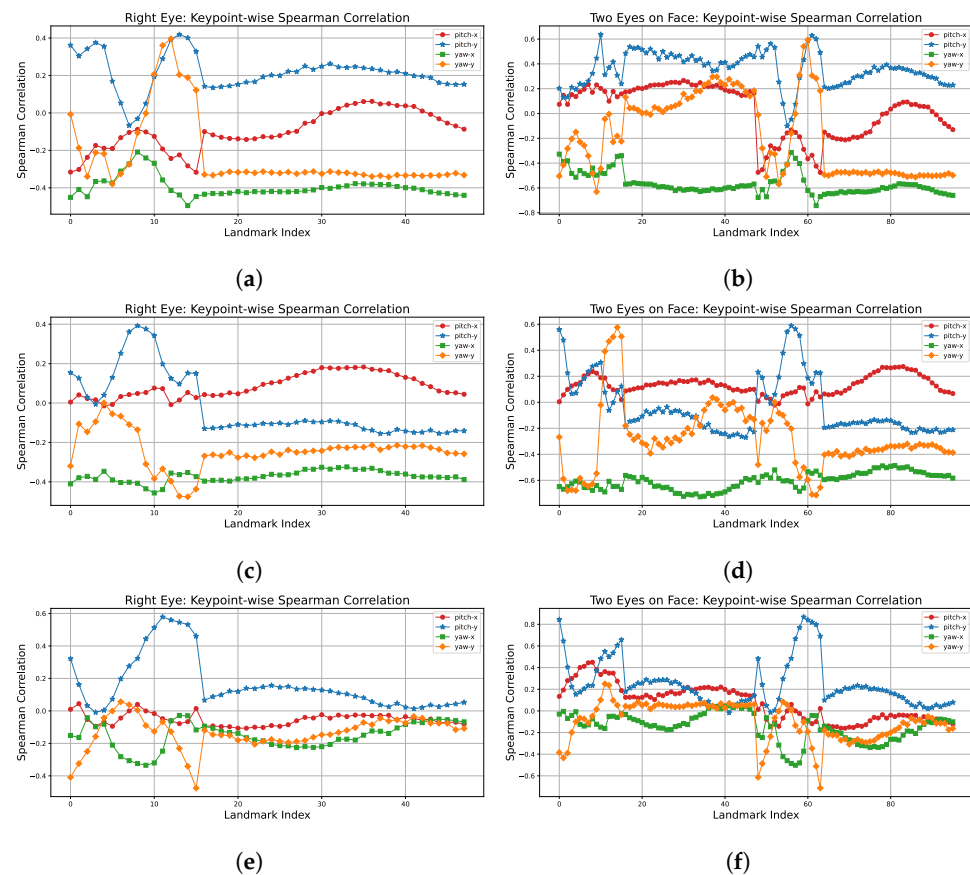


Figure 4. Spearman correlation between keypoints and gaze on MPIIFaceGaze. (a) Right-eye single-eye analysis for the Person 0; (b) Dual-eye analysis combining keypoints from both eyes for the Person 0; (c) Right-eye single-eye analysis for an additional subject (Person 1); (d) Dual-eye analysis for Person 1; (e) Right-eye single-eye analysis for another subject (Person 2); (f) Dual-eye analysis for Person 2.

In addition, to verify that the observed correlation patterns are not specific to a single individual, we repeat the same Spearman correlation analysis on two additional subjects from the MPIIFaceGaze validation set. The results are shown in Figure 4c–f, where Figure 4c,d and Figure 4e,f correspond to two other subjects, respectively, under the same single-eye and dual-eye settings as in Figure 4a,b. Despite individual variations in correlation magnitude, all three subjects exhibit consistent and interpretable trends: pitch remains more strongly correlated with the y-coordinates of eye keypoints, while yaw shows higher correlation with x-coordinates, particularly in subsets of keypoints associated with horizontal eye motion. In the dual-eye setting, keypoints from both eyes consistently contribute to pitch-related correlations, whereas yaw-related correlations tend to be more concentrated in specific eye regions.

These subject-wise consistent patterns indicate that the relationship between predicted eye keypoints and gaze direction is stable across different individuals, rather than an artifact of a single example. This further supports the robustness and interpretability of the proposed keypoint-based representation for gaze estimation.

4.8. Robustness Evaluation and Visualization

The robustness of the proposed method was evaluated using subjects wearing glasses from the MPIIFaceGaze dataset, with the results summarized in Table 5. A comparison be-

tween the baseline method and LGNet on these subjects highlights the superior robustness of our approach, particularly in handling occlusion within the eye region. Additionally, Figure 5 presents visualizations of the keypoints and 3D gaze estimates centered on the face for several glasses-wearing subjects from the MPIIFaceGaze dataset, further illustrating the effectiveness of LGNet in maintaining accurate gaze estimation despite the occlusions caused by eyewear.

MPIIFaceGaze

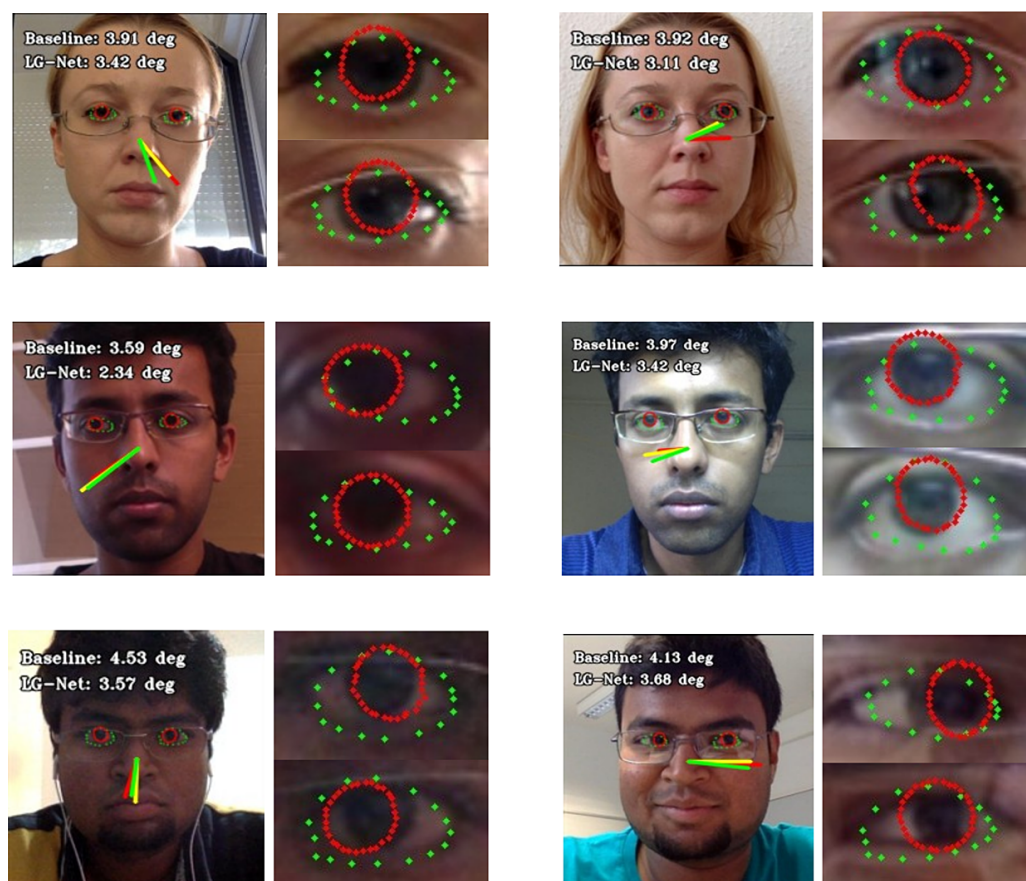


Figure 5. Visualization of keypoints predictions and 3D gaze for glasses-wearing subjects (images sourced from the MPIIFaceGaze dataset [8]). Red, green, and yellow lines denote the ground-truth gaze vector, baseline prediction, and LGNet prediction, respectively.

Table 5. Angle error on glasses-wearing subjects in MPIIFaceGaze, based on subjects P01, P04, P07, P10, and P11.

Model	MPIIFaceGaze
Baseline	4.06
+CVAE	3.77
+CVAE+SSFF	3.51
LGNet	3.46

5. Discussion

Despite the promising performance of LGNet, several limitations and challenges remain. First, the eye-keypoint generation network is pretrained on the synthetic UnityEyes dataset, which may introduce a domain gap when applied to real-world images under extreme illumination, heavy occlusion, or uncommon eye appearances. Although experi-

mental results demonstrate good generalization, further improvements could be achieved by incorporating more diverse real-world annotations or domain adaptation strategies. Second, the proposed framework introduces additional computational complexity due to the dual-eye CVAE-based keypoint generators, multi-scale feature fusion, and cross-attention modules. While this design improves accuracy and robustness, it may lead to increased inference time compared to lightweight gaze estimation models, limiting deployment on resource-constrained devices. Third, although LGNet performs well under moderately unconstrained conditions, its robustness has not been explicitly evaluated under certain challenging visual scenarios, such as extreme head poses or severe motion blur. In addition, the current framework does not explicitly model temporal information, which may further affect performance under dynamic conditions.

Future work will focus on improving computational efficiency, enhancing robustness under extreme conditions, and exploring lightweight or temporal extensions to further improve scalability and real-world applicability.

6. Conclusions

This paper presents LGNet, a novel approach for 3D gaze estimation that integrates geometric priors with visual features. To address the limitations of existing methods, particularly in challenging scenarios such as occlusion, LGNet introduces a keypoint generation network coupled with a multi-scale feature fusion mechanism. The method leverages a CVAE, trained on synthetic data, to generate eye keypoints. The pretrained prior network is then applied to gaze estimation datasets, enabling effective geometric modeling. To achieve a deeper fusion of keypoints with facial features, we design an SSFF module and incorporate a dual-stream cross-attention mechanism. Additionally, a GCR module is introduced to guide the model in focusing on the most informative regions during feature fusion. Experimental results demonstrate that LGNet outperforms current state-of-the-art approaches across multiple datasets, achieving superior accuracy, particularly in occluded conditions. However, the use of dual-eye-keypoint generators and a facial feature extractor introduces increased model complexity, which may slow down inference speed and reduce effectiveness on low-resolution images. Future work will focus on strategies to enhance performance in low-resolution settings and improve model efficiency, aiming to develop a more lightweight and scalable 3D gaze estimation solution.

Author Contributions: Conceptualization, Y.L., D.L. and G.X.; Methodology, Y.L., D.L., X.Z., W.Z. and G.X.; Software, X.Z. and W.Z.; Validation, X.Z. and W.Z.; Formal analysis, Z.M. and L.D.; Investigation, Z.M. and L.D.; Resources, Y.L.; Data curation, X.Z. and W.Z.; Writing—original draft preparation, X.Z., Y.L. and D.L.; Writing—review and editing, Y.L., D.L., Z.M., L.D. and G.X.; Visualization, X.Z., Z.M. and W.Z.; Supervision, D.L., Y.L. and G.X.; Project administration, D.L. and Y.L.; Funding acquisition, D.L. and L.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (Nos. 62507035, 62307034).

Data Availability Statement: The code supporting the findings of this study is publicly available at: <https://github.com/dacilab/LGNet> (accessed on 9 February 2026).

Conflicts of Interest: Author Zhongjie Mao was employed by the company Wuhan HomeLightyear Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Steil, J.; Hagedstedt, I.; Huang, M.X.; Bulling, A. Privacy-aware eye tracking using differential privacy. In Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, Denver, CO, USA, 25–28 June 2019; pp. 1–9.

2. Wang, S.; Zhao, A.; Lai, C.; Zhang, Q.; Li, D.; Gao, Y.; Dong, L.; Wang, X. Gcanet: Geometry cues-aware facial expression recognition based on graph convolutional networks. *J. King Saud-Univ.-Comput. Inf. Sci.* **2023**, *35*, 101605. [CrossRef]
3. Zhao, W.; Wang, S.; Wang, X.; Li, D.; Wang, J.; Lai, C.; Li, X. Dadl: Double asymmetric distribution learning for head pose estimation in wisdom museum. *J. King Saud-Univ.-Comput. Inf. Sci.* **2024**, *36*, 101869. [CrossRef]
4. Liu, H.; Fang, S.; Zhang, Z.; Li, D.; Lin, K.; Wang, J. MFDNet: Collaborative poses perception and matrix Fisher distribution for head pose estimation. *IEEE Trans. Multimed.* **2021**, *24*, 2449–2460. [CrossRef]
5. Wu, X.; Li, L.; Zhou, G.; Wu, Q.; Zuo, X.; Zhu, H.; He, S. Multi-task driver gaze estimation in real world driving scenes. *Eng. Appl. Artif. Intell.* **2025**, *160*, 111892. [CrossRef]
6. Wood, E.; Baltrušaitis, T.; Morency, L.-P.; Robinson, P.; Bulling, A. Learning an appearance-based gaze estimator from one million synthesised images. In Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, Charleston, SC, USA, 14–17 March 2016; pp. 131–138.
7. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 162–175. [CrossRef] [PubMed]
8. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 51–60.
9. 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 Proceedings of the Symposium on Eye Tracking Research and Applications, Safety Harbor, FL, USA, 26–28 March 2014; pp. 255–258.
10. Kellnhofer, P.; Recasens, A.; Stent, S.; Matusik, W.; Torralba, A. Gaze360: Physically unconstrained gaze estimation in the wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6912–6921.
11. 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 *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020; Proceedings, Part V 16, 2020; pp. 365–381.
12. Alberto Funes Mora, K.; Odobez, J.-M. Geometric generative gaze estimation (g3e) for remote rgb-d cameras. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1773–1780.
13. Cheng, Y.; Lu, F.; Zhang, X. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 100–115.
14. Park, S.; Spurr, A.; Hilliges, O. Deep pictorial gaze estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 721–738.
15. Park, S.; Zhang, X.; Bulling, A.; Hilliges, O. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, Warsaw, Poland, 14–17 June 2018; pp. 1–10.
16. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. Appearance-based gaze estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4511–4520.
17. Huang, G.; Shi, J.; Xu, J.; Li, J.; Chen, S.; Du, Y.; Zhen, X.; Liu, H. Gaze estimation by attention-induced hierarchical variational auto-encoder. *IEEE Trans. Cybern.* **2023**, *54*, 2592–2605. [CrossRef] [PubMed]
18. Li, D.; Wang, S.; Zhao, W.; Kang, L.; Dong, L.; Wang, J.; Wang, X. Adgaze: Anisotropic gaussian label distribution learning for fine-grained gaze estimation. *Pattern Recognit.* **2025**, *164*, 111536. [CrossRef]
19. Gou, C.; Yu, Y.; Guo, Z.; Xiong, C.; Cai, M. Cascaded learning with transformer for simultaneous eye landmark, eye state and gaze estimation. *Pattern Recognit.* **2024**, *156*, 110760. [CrossRef]
20. Sohn, K.; Lee, H.; Yan, X. Learning structured output representation using deep conditional generative models. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 3483–3491.
21. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114. Available online: <https://arxiv.org/abs/1312.6114> (accessed on 9 February 2026).
22. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Springer: Cham, Switzerland, 2015; Proceedings, Part III 18; pp. 234–241.
23. Zheng, C.; Mendieta, M.; Chen, C. Poster: A pyramid cross-fusion transformer network for facial expression recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 3146–3155.
24. Kingma, D.P. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
25. Paszke, A. Pytorch: An imperative style, high-performance deep learning library. *arXiv* **2019**, arXiv:1912.01703.
26. Chen, Z.; Shi, B.E. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 309–324.

27. Cheng, Y.; Huang, S.; Wang, F.; Qian, C.; Lu, F. A coarse-to-fine adaptive network for appearance-based gaze estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10623–10630.
28. Wu, C.; Hu, H.; Lin, K.; Wang, Q.; Liu, T.; Chen, G. Attention-guided and fine-grained feature extraction from face images for gaze estimation. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106994. [[CrossRef](#)]
29. Cheng, Y.; Lu, F. Gaze estimation using transformer. In *Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022*; IEEE: New York, NY, USA, 2022; pp. 3341–3347.
30. Shi, Y.; Zhang, F.; Yang, W.; Wang, G.; Su, N. Agent-guided gaze estimation network by two-eye asymmetry exploration. In *Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 27–30 October 2024*; IEEE: New York, NY, USA, 2024; pp. 2320–2326.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.